(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 722 165 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
17.07.1996 Bulletin 1996/29

(51) Int Cl.[6]: **G10L 9/14**

(21) Application number: **96300245.6**

(22) Date of filing: **12.01.1996**

(84) Designated Contracting States:
DE FR GB SE

(30) Priority: **12.01.1995 US 371743**

(71) Applicant: **DIGITAL VOICE SYSTEMS, INC.**
Burlington,MA 01803 (US)

(72) Inventor: **Griffin, Daniel Wayne**
Hollis, New Hampshire 03049 (US)

(74) Representative: **Deans, Michael John Percy et al**
Lloyd Wise, Tregear & Co.,
Commonwealth House,
1-19 New Oxford Street
London WC1A 1LW (GB)

(54) Estimation of excitation parameters

(57) Excitation parameters for a digitized speech signal are determined by analysing the digitized speech signal. The digitized speech signal is divided into at least two frequency bands. A first preliminary excitation parameter is determined by performing a nonlinear operation on at least one of the frequency band signals to produce a modified frequency band signal and determining the first preliminary excitation parameter using the modified frequency band signal. A second preliminary excitation parameter is determined using a method different from the first method. The first and second preliminary excitation parameters are used to determine an excitation parameter for the digitized speech signal. The method is useful in encoding speech. Speech synthesized using the parameters estimated based on the invention generates high quality speech at various bit rates useful for applications such as satellite voice communication.

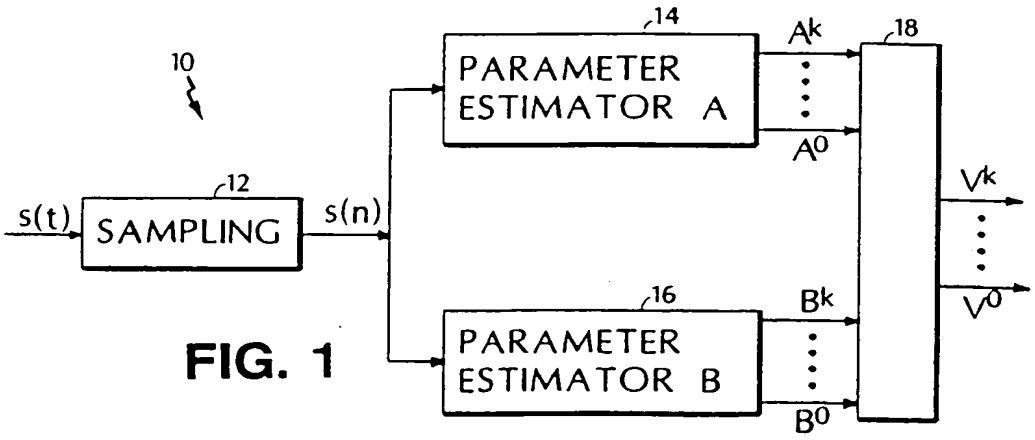

FIG. 1

EP 0 722 165 A2 CORRIGENDUM issued on 20.11.1996 (bibliography updates included)

Description

The invention has arisen from work seeking to improve the accuracy with which excitation parameters are estimated in speech analysis and synthesis.

Speech analysis and synthesis are widely used in applications such as telecommunications and voice recognition. A vocoder, which is a type of speech analysis/synthesis system, models speech as the response of a system to excitation over short time intervals. Examples of vocoder systems include linear prediction vocoders, homomorphic vocoders, channel vocoders, sinusoidal transform coders ("STC"), multiband excitation ("MBE") vocoders, improved multiband excitation ("IMBE (TM)") vocoders.

Vocoders typically synthesize speech based on excitation parameters and system parameters. Typically, an input signal is segmented using, for example, a Hamming window. Then, for each segment, system parameters and excitation parameters are determined. System parameters include the spectral envelope or the impulse response of the system. Excitation parameters include a fundamental frequency (or pitch) and a voiced/unvoiced parameter that indicates whether the input signal has pitch (or indicates the degree to which the input signal has pitch). In vocoders that divide the speech into frequency bands, such as IMBE (TM) vocoders, the excitation parameters may also include a voiced/unvoiced parameter for each frequency band rather than a single voiced/unvoiced parameter. Accurate excitation parameters are essential for high quality speech synthesis.

When the voiced/unvoiced parameters include only a single voiced/unvoiced decision for the entire frequency band, the synthesized speech tends to have a "buzzy" quality especially noticeable in regions of speech which contain mixed voicing or in voiced regions of noisy speech. A number of mixed excitation models have been proposed as potential solutions to the problem of "buzziness" in vocoders. In these models, periodic and noise-like excitations are mixed which have either time-invariant or time-varying spectral shapes.

In excitation models having time-invariant spectral shapes, the excitation signal consists of the sum of a periodic source and a noise source with fixed spectral envelopes. The mixture ratio controls the relative amplitudes of the periodic and noise sources. Examples of such models include Itakura and Saito, "Analysis Synthesis Telephony Based upon the Maximum Likelihood Method," Reports of 6th Int. Cong. Acoust., Tokyo, Japan, Paper C-5-5, pp. C17-20, 1968; and Kwon and Goldberg, "An Enhanced LPC Vocoder with No Voiced/Unvoiced Switch," IEEE Trans. on Acoust., Speech, and Signal Processing, vol. ASSP-32, no. 4, pp. 851-858, August 1984. In theses excitation models a white noise source is added to a white periodic source. The mixture ratio between these sources is estimated from the height of the peak of the autocorrelation of the LPC residual.

In excitation models having time-varying spectral shapes, the excitation signal consists of the sum of a periodic source and a noise source with time varying spectral envelope shapes. Examples of such models include Fujimara, "An Approximation to Voice Aperiodicity," IEEE Trans. Audio and Electroacoust., pp. 68-72, March 1968; Makhoul et al., "A Mixed-Source Excitation Model for Speech Compression and Synthesis," IEEE Int. Conf. on Acoust. Sp. & Sig. Proc., April 1978, pp. 163-166; Kwon and Goldberg, "An Enhanced LPC Vocoder with No Voiced/Unvoiced Switch," IEEE Trans. on Acoust., Speech, and Signal Processing, vol. ASSP-32, no. 4, pp. 851-858, August 1984; and Griffin and Lim, "Multiband Excitation Vocoder," IEEE Trans.
Acoust., Speech, Signal Processing, vol. ASSP-36, pp. 1223-1235, Aug. 1988.

In the excitation model proposed by Fujimara, the excitation spectrum is divided into three fixed frequency bands. A separate cepstral analysis is performed for each frequency band and a voiced/unvoiced decision for each frequency band is made based on the height of the cepstrum peak as a measure of periodicity.

In the excitation model proposed by Makhoul et al., the excitation signal consists of the sum of a low-pass periodic source and a high-pass noise source. The low-pass periodic source is generated by filtering a white pulse source with a variable cut-off low-pass filter. Similarly, the high-pass noise source was generated by filtering a white noise source with a variable cut-off high-pass filter. The cut-off frequencies for the two filters are equal and are estimated by choosing the highest frequency at which the spectrum is periodic. Periodicity of the spectrum is determined by examining the separation between consecutive peaks and determining whether the separations are the same, within some tolerance level.

In a second excitation model implemented by Kwon and Goldberg, a pulse source is passed through a variable gain low-pass filter and added to itself, and a white noise source is passed through a variable gain high-pass filter and added to itself. The excitation signal is the sum of the resultant pulse and noise sources with the relative amplitudes controlled by a voiced/unvoiced mixture ratio. The filter gains and voiced/unvoiced mixture ratio are estimated from the LPC residual signal with the constraint that the spectral envelope of the resultant excitation signal is flat.

In the multiband excitation model proposed by Griffin and Lim, a frequency dependent voiced/unvoiced mixture function is proposed. This model is restricted to a frequency dependent binary voiced/unvoiced decision for coding purposes. A further restriction of this model divides the spectrum into a finite number of frequency bands with a binary voiced/unvoiced decision for each band. The voiced/unvoiced information is estimated by comparing the speech spectrum to the closest periodic spectrum. When the error is below a threshold, the band is marked voiced, otherwise, the

band is marked unvoiced.

Excitation parameters may also be used in applications, such as speech recognition, where no speech synthesis is required. Once again, the accuracy of the excitation parameters directly affects the performance of such a system.

In one aspect, generally, the invention features a hybrid excitation parameter estimation technique that produces two sets of excitation parameters for a speech signal using two different approaches and combines the two sets to produce a single set of excitation parameters. In a first approach, the technique applies a nonlinear operation to the speech signal to emphasize the fundamental frequency of the speech signal. In a second approach, we use a different method which may or may not include a nonlinear operation. While the first approach produces highly accurate excitation parameters under most conditions, the second approach produces more accurate parameters under certain conditions. By using both approaches and combining the resulting sets of excitation parameters to produce a single set, our technique produces accurate results under a wider range of conditions than are produced by either of the approaches individually.

In typical approaches to determining excitation parameters, an analog speech signal s(t) is sampled to produce a speech signal s(n). Speech signal s(n) is then multiplied by a window w(n) to produce a windowed signal $s_w(n)$ that is commonly referred to as a speech segment or a speech frame. A Fourier transform is then performed on windowed signal $s_w(n)$ to produce a frequency spectrum $S_w(\omega)$ from which the excitation parameters are determined.

When speech signal s(n) is periodic with a fundamental frequency ω_o or pitch period n_o (where n_o equals $2\pi/\omega_o$), the frequency spectrum of speech signal s(n) should be a line spectrum with energy at ω_o and harmonics thereof (integral multiples of ω_o). As expected, $S_w(\omega)$ has spectral peaks that are centered around ω_o and its harmonics. However, due to the windowing operation, the spectral peaks include some width, where the width depends on the length and shape of window w(n) and tends to decrease as the length of window w(n) increases. This window-induced error reduces the accuracy of the excitation parameters. Thus, to decrease the width of the spectral peaks, and to thereby increase the accuracy of the excitation parameters, the length of window w(n) should be made as long as possible.

The maximum useful length of window w(n) is limited. Speech signals are not stationary signals, and instead have fundamental frequencies that change over time. To obtain meaningful excitation parameters, an analyzed speech segment must have a substantially unchanged fundamental frequency. Thus, the length of window w(n) must be short enough to ensure that the fundamental frequency will not change significantly within the window.

In addition to limiting the maximum length of window w(n), a changing fundamental frequency tends to broaden the spectral peaks. This broadening effect increases with increasing frequency. For example, if the fundamental frequency changes by $\Delta\omega_o$ during the window, the frequency of the *mth* harmonic, which has a frequency of $m\omega_o$, changes by $m\Delta\omega_o$ so that the spectral peak corresponding to $m\omega_o$ is broadened more than the spectral peak corresponding to ω_o. This increased broadening of the higher harmonics reduces the effectiveness of higher harmonics in the estimation of the fundamental frequency and the generation of voiced/unvoiced parameters for high frequency bands.

By applying a nonlinear operation to the speech signal, the increased impact on higher harmonics of a changing fundamental frequency is reduced or eliminated, and higher harmonics perform better in estimation of the fundamental frequency and determination of voiced/unvoiced parameters. Suitable nonlinear operations map from complex (or real) to real values and produce outputs that are nondecreasing functions of the magnitudes of the complex (or real) values. Such operations include, for example, the absolute value, the absolute value squared, the absolute value raised to some other power, or the log of the absolute value.

Nonlinear operations tend to produce output signals having spectral peaks at the fundamental frequencies of their input signals. This is true even when an input signal does not have a spectral peak at the fundamental frequency. For example, if a bandpass filter that only passes frequencies in the range between the third and fifth harmonics of ω_o is applied to a speech signal s(n), the output of the bandpass filter, x(n), will have spectral peaks at $3\omega_o$, $4\omega_o$ and $5\omega_o$.

Though x(n) does not have a spectral peak at ω_o, $|x(n)|^2$ will have such a peak. For a real signal x(n), $|x(n)|^2$ is equivalent to $x^2(n)$. As is well known, the Fourier transform of $x^2(n)$ is the convolution of $X(\omega)$, the Fourier transform of x(n), with $X(\omega)$:

$$\sum_{n=-\infty}^{\infty} x^2(n)\, e^{-j\omega n} = \frac{1}{2\pi} \int_{u=-\pi}^{\pi} X(\omega-u)\, X(u)\, du.$$

The convolution of $X(\omega)$ with $X(\omega)$ has spectral peaks at frequencies equal to the differences between the frequencies for which $X(\omega)$ has spectral peaks. The differences between the spectral peaks of a periodic signal are the fundamental frequency and its multiples. Thus, in the example in which $X(\omega)$ has spectral peaks at $3\omega_o$, $4\omega_o$ and $5\omega_o$, $X(\omega)$ convolved with $X(\omega)$ has a spectral peak at ω_o ($4\omega_o$-$3\omega_o$, $5\omega_o$-$4\omega_o$). For a typical periodic signal, the spectral peak at the funda-

mental frequency is likely to be the most prominent.

The above discussion also applies to complex signals. For a complex signal x(n), the Fourier transform of $|x(n)|^2$ is:

$$\sum_{n=-\infty}^{\infty} |x(n)|^2 e^{-j\omega n} = \frac{1}{2\pi} \int_{u=-\pi}^{\pi} X(\omega+u)\, X^*(u)\, du.$$

This is an autocorrelation of $X(\omega)$ with $X^*(\omega)$, and also has the property that spectral peaks separated by $n\omega_o$ produce peaks at $n\omega_o$.

Even though $|x(n)|$, $|x(n)|^a$ for some real "a", and $\log |x(n)|$ are not the same as $|x(n)|^2$, the discussion above for $|x(n)|^2$ applies approximately at the qualitative level. For example, for $|x(n)| = y(n)^{0.5}$, where $y(n) = |x(n)|^2$, a Taylor series expansion of y(n) can be expressed as:

$$|x(n)| = \sum_{k=0}^{\infty} c_k y^k(n) .$$

Because multiplication is associative, the Fourier transform of the signal $y^k(n)$ is $Y(\omega)$ convolved with the Fourier transform of $y^{k-1}(n)$. The behavior for nonlinear operations other than $|x(n)|^2$ can be derived from $|x(n)|^2$ by observing the behavior of multiple convolutions of $Y(\omega)$ with itself. If $Y(\omega)$ has peaks at $n\omega_o$, then multiple convolutions of $Y(\omega)$ with itself will also have peaks at $n\omega_o$.

As shown, nonlinear operations emphasize the fundamental frequency of a periodic signal, and are particularly useful when the periodic signal includes significant energy at higher harmonics. However, the presence of the nonlinearity can degrade performance in some cases. For example, performance may be degraded when speech signal s(n) is divided into multiple bands $s^i(n)$ using bandpass filters, where $s^i(n)$ denotes the result of bandpass filtering using the ith bandpass filter. If a single harmonic of the fundamental frequency is present in the pass band of the ith filter, the output of the filter is:

$$s^i(n) = A_k e^{j(\omega_k n + \theta_k)}$$

where ω_k is the frequency, θ_k is the phase, and A_k is the amplitude of the harmonic. When a nonlinearity such as the absolute value is applied to $s^i(n)$ to produce a value $y^i(n)$, the result is:

$$y^i(n) = |s^i(n)| = |A_k|$$

so that the frequency information has been completely removed from the signal $y^i(n)$. Removal of this frequency information can reduce the accuracy of parameter estimates.

Our hybrid technique provides significantly improved parameter estimation performance in cases for which the nonlinearity reduces the accuracy of parameter estimates while maintaining the benefits of the nonlinearity in the remaining cases. As described above, the hybrid technique includes combining parameter estimates based on the signal after the nonlinearity has been applied ($y^i(n)$) with parameter estimates based on the signal before the nonlinearity is applied ($s^i(n)$ or s(n)). The two approaches produce parameter estimates along with an indication of the probability of correctness of these parameter estimates. The parameter estimates are then combined giving higher weight to estimates with a higher probability of being correct.

In another aspect, generally, the invention features the application of smoothing techniques to the voiced/unvoiced parameters. Voiced/unvoiced parameters can be binary or continuous functions of time and/or frequency. Because these parameters tend to be smooth functions in at least one direction (positive or negative) of time or frequency, the estimates of these parameters can benefit from appropriate application of smoothing techniques in time and/or frequency.

The invention also features an improved technique for estimating voiced/unvoiced parameters. In vocoders such as linear prediction vocoders, homomorphic vocoders, channel vocoders, sinusoidal transform coders, multiband excitation vocoders, and IMBE (TM) vocoders, a pitch period n (or equivalently a fundamental frequency) is selected.

Thereafter, a function $f^i(n)$ is then evaluated at the selected pitch period (or fundamental frequency) to estimate the ith voiced/unvoiced parameter. However, for some speech signals, evaluation of this function only at the selected pitch period will result in reduced accuracy of one or more voiced/unvoiced parameter estimates. This reduced accuracy may result from speech signals that are more periodic at a multiple of the pitch period than at the pitch period, and may be frequency dependent so that only certain portions of the spectrum are more periodic at a multiple of the pitch period. Consequently, the voiced/unvoiced parameter estimation accuracy can be improved by evaluating the function $f^i(n)$ at the pitch period n and at its multiples, and thereafter combining the results of these evaluations.

In another aspect, the invention features an improved technique for estimating the fundamental frequency or pitch period. When the fundamental frequency ω_o (or pitch period n_o) is estimated, there may be some ambiguity as to whether ω_o or a submultiple or multiple of ω_o is the best choice for the fundamental frequency. Since the fundamental frequency tends to be a smooth function of time for voiced speech, predictions of the fundamental frequency based on past estimates can be used to resolve ambiguities and improve the fundamental frequency estimate.

Other features and advantages of the invention will be apparent from the following description of the preferred embodiments, given by way of example only.

In the drawings:

Fig. 1 is a block diagram of a system for determining whether frequency bands of a signal are voiced or unvoiced.

Fig. 2 is a block diagram of a parameter estimation unit of the system of Fig. 1.

Fig. 3 is a block diagram of a channel processing unit of the parameter estimation unit of Fig. 2.

Fig. 4 is a block diagram of a parameter estimation unit of the system of Fig. 1.

Fig. 5 is a block diagram of a channel processing unit of the parameter estimation unit of Fig. 4.

Fig. 6 is a block diagram of a parameter estimation unit of the system of Fig. 1.

Fig. 7 is a block diagram of a channel processing unit of the parameter estimation unit of Fig. 6.

Figs. 8-10 are block diagrams of systems for determining the fundamental frequency of a signal.

Fig. 11 is a block diagram of voiced/unvoiced parameter smoothing unit.

Fig. 12 is a block diagram of voiced/unvoiced parameter improvement unit.

Fig. 13 is a block diagram of a fundamental frequency improvement unit.

Figs. 1-12 show the structure of a system for estimating excitation parameters, the various blocks and units of which are preferably implemented with software.

With reference to Fig. 1, a voiced/unvoiced determination system 10 includes a sampling unit 12 that samples an analog speech signal s(t) to produce a speech signal s(n). For typical speech coding applications, the sampling rate ranges between six kilohertz and ten kilohertz.

Speech signal s(n) is supplied to a first parameter estimator 14 that divides the speech signal into k+1 bands and produces a first set of preliminary voiced/unvoiced ("V/UV") parameters (A^0 to A^K) corresponding to a first estimate as to whether the signals in the bands are voiced or unvoiced. Speech signal s(n) is also supplied to a second parameter estimator 16 that produces a second set of preliminary V/UV parameters (B^0 to B^K) that correspond to a second estimate as to whether the signals in the bands are voiced or unvoiced. The two sets of preliminary V/UV parameters are combined by a combination block 18 to produce a set of V/UV parameters (V^0 to V^K).

With reference to Fig. 2, first parameter estimator 14 produces the first voiced/unvoiced estimate using a frequency domain approach. Channel processing units 20 in first parameter estimator 14 divide speech signal s(n) into at least two frequency bands and process the frequency bands to produce a first set of frequency band signals, designated as $T^0(\omega)$ .. $T^I(\omega)$. As discussed below, channel processing units 20 are differentiated by the parameters of a bandpass filter used in the first stage of each channel processing unit 20. In the described embodiment, there are sixteen channel processing units (I equals 15).

A remap unit 22 transforms the first set of frequency band signals to produce a second set of frequency band signals, designated as $U^0(\omega)$ .. $U^K(\omega)$. In the described embodiment, there are eight frequency band signals in the second set of frequency band signals (K equals 7). Thus, remap unit 22 maps the frequency band signals from the sixteen channel processing units 20 into eight frequency band signals. Remap unit 20 does so by combining consecutive pairs of frequency band signals from the first set into single frequency band signals in the second set. For example, $T^0(\omega)$ and $T^1(\omega)$ are combined to produce $U^0(\omega)$, and $T^{14}(\omega)$ and $T^{15}(\omega)$ are combined to produce $U^7(\omega)$. Other approaches to remapping could also be used.

Next, voiced/unvoiced parameter estimation units 24, each associated with a frequency band signal from the second set, produce preliminary V/UV parameters A^0 to A^K by computing a ratio of the voiced energy in the frequency band at an estimated fundamental frequency ω^o to the total energy in the frequency band and subtracting this ratio from 1:

$$A^k = 1.0 - E_v^k(\omega_o) / E_t^k .$$

The voiced energy in the frequency band is computed as:

$$E_V^k(\omega_o) = \sum_{n=1}^{N} \sum_{\omega_m \in I_n} U^k(\omega_m)$$

where

$$I_n = [(n\text{-}0.25)\ \omega_o,\ (n\text{+}0.25)\ \omega_o],$$

and N is the number of harmonics of the fundamental frequency ω_o being considered. V/UV parameter estimation units 24 determine the total energy of their associated frequency band signals as:

$$E_T^k = \sum_{\forall \omega_m > 0.5\omega_o} U^k(\omega_m) .$$

The degree to which the frequency band signal is voiced varies indirectly with the value of the preliminary V/UV parameter. Thus, the frequency band signal is highly voiced when the preliminary V/UV parameter is near zero and is highly unvoiced when the parameter is greater than or equal to one half.

With reference to Fig. 3, when speech signal s(n) enters a channel processing unit 20, components $s^i(n)$ belonging to a particular frequency band are isolated by a bandpass filter 26. Bandpass filter 26 uses downsampling to reduce computational requirements, and does so without any significant impact on system performance. Bandpass filter 26 can be implemented as a Finite Impulse Response (FIR) or Infinite Impulse Response (IIR) filter, or by using an FFT. In the described embodiment, bandpass filter 26 is implemented using a thirty two point real input FFT to compute the outputs of a thirty two point FIR filter at seventeen frequencies, and achieves a downsampling factor of S by shifting the input by S samples each time the FFT is computed. For example, if a first FFT used samples one through thirty two, a downsampling factor of ten would be achieved by using samples eleven through forty two in a second FFT.

A first nonlinear operation unit 28 then performs a nonlinear operation on the isolated frequency band $s^i(n)$ to emphasize the fundamental frequency of the isolated frequency band $s^i(n)$. For complex values of $s^i(n)$ (i greater than zero), the absolute value, $|s^i(n)|$, is used. For the real value of $s^0(n)$, $s^0(n)$ is used if $s^0(n)$ is greater than zero and zero is used if $s^0(n)$ is less than or equal to zero.

The output of nonlinear operation unit 28 is passed through a lowpass filtering and downsampling unit 30 to reduce the data rate and consequently reduce the computational requirements of later components of the system. Lowpass filtering and downsampling unit 30 uses an FIR filter computed every other sample for a downsampling factor of two.

A windowing and FFT unit 32 multiplies the output of lowpass filtering and downsampling unit 30 by a window and computes a real input FFT, $S^i(\omega)$, of the product. Typically, windowing and FFT unit 32 uses a Hamming window and a real input FFT.

Finally, a second nonlinear operation unit 34 performs a nonlinear operation on $S^i(\omega)$ to facilitate estimation of voiced or total energy and to ensure that the outputs of channel processing units 20, $T^i(\omega)$, combine constructively if used in fundamental frequency estimation. The absolute value squared is used because it makes all components of $T^i(\omega)$ real and positive.

With reference to Fig. 4, second parameter estimator 16 produces the second preliminary V/UV estimates using a sinusoid detector/estimator. Channel processing units 36 in second parameter estimator 16 divide speech signal s(n) into at least two frequency bands and process the frequency bands to produce a first set of signals, designated as $R^0(I)$ .. $R^I(I)$. Channel processing units 36 are differentiated by the parameters of a bandpass filter used in the first stage of each channel processing unit 36. In the described embodiment, there are sixteen channel processing units (I equals 15). The number of channels (value of I) in Fig. 4 does not have to equal the number of channels (value of I) in Fig. 2.

A remap unit 38 transforms the first set of signals to produce a second set of signals, designated as $S^0(I)$ .. $S^K(I)$. The remap unit can be an identity system. In the described embodiment, there are eight signals in the second set of signals (K equals 7). Thus, remap unit 38 maps the signals from the sixteen channel processing units 36 into eight signals. Remap unit 38 does so by combining consecutive pairs of signals from the first set into single signals in the

second set. For example, $R^0(l)$ and $R^1(l)$ are combined to produce $S^0(l)$, and $R^{14}(l)$ and $R^{15}(l)$ are combined to produce $S^7(l)$. Other approaches to remapping could also be used.

Next, V/UV parameter estimation units 40, each associated with a signal from the second set, produce preliminary V/UV parameters B^0 to B^K by computing a ratio of the sinusoidal energy in the signal to the total energy in the signal and subtracting this ratio from 1:

$$B^k = 1.0 - S^k(1)/S^k(0).$$

With reference to Fig. 5, when speech signal s(n) enters a channel processing unit 36, components $s^i(n)$ belonging to a particular frequency band are isolated by a bandpass filter 26 that operates identically to the bandpass filters of channel processing units 20 (see Fig. 3). It should be noted that, to reduce computation requirements, the same bandpass filters may be used in channel processing units 20 and 36, with the outputs of each filter being supplied to a first nonlinear operation unit 28 of a channel processing unit 20 and a window and correlate unit 42 of a channel processing unit 36.

A window and correlate unit 42 then produces two correlation values for the isolated frequency band $s^i(n)$. The first value, $R^i(0)$, provides a measure of the total energy in the frequency band:

$$R^i(0) = \left[\frac{1}{2}\sum_{n=0}^{N-1}\left[\,|s^i(n)|^2 + |s^i(n+S)|^2\right]\right]^2$$

where N is related to the size of the window and typically defines an interval of 20 milliseconds and S is the number of samples by which the bandpass filter shifts the input speech samples. The second value, $R^i(1)$, provides a measure of the sinusoidal energy in the frequency band:

$$R^i(1) = \left|\sum_{n=0}^{N-1} s^i(n+S)\, s^{*i}(n)\right|^2.$$

Combination block 18 produces voiced/unvoiced parameters V^0 to V^K by selecting the minimum of a preliminary V/UV parameter from the first set and a function of a preliminary V/UV parameter from the second set. In particular, combination block produces the voiced/unvoiced parameters as:

$$V^k = \min(A^k, f_B(B^k))$$

where

$$f_B(B^k) = B^k + \alpha(k)\,\beta(\omega_o),$$

$$\beta(\omega_o) = 1.0, \text{ when } \omega_o \geq 2\pi/60.0, \text{ or}$$

$$2\pi/(60\omega_o), \text{ when } \omega_o < 2\pi/60.0$$

and $\alpha(k)$ is an increasing function of k. Because a preliminary V/UV parameter having a value close to zero has a higher probability of being correct than a preliminary V/UV parameter having a larger value, the selection of the minimum value results in the selection of the value that is most likely to be correct.

With reference to Fig. 6, in another embodiment, a first parameter estimator 14' produces the first preliminary V/UV estimate using an autocorrelation domain approach. Channel processing units 44 in first parameter estimator 14' divide speech signal s(n) into at least two frequency bands and process the frequency bands to produce a first set of

frequency band signals, designated as $T^0(l)$ .. $T^K(l)$. There are eight channel processing units (K equals 7) and no remapping unit is necessary.

Next, voiced/unvoiced (V/UV) parameter estimation units 46, each associated with a channel processing unit 44, produce preliminary V/UV parameters A^0 to A^K by computing a ratio of the voiced energy in the frequency band at an estimated pitch period n_o to the total energy in the frequency band and subtracting this ratio from 1:

$$A^k = 1.0 - E_v^k(n_o) / E_t^k .$$

The voiced energy in the frequency band is computed as:

$$E_v^k(n_o) = C(n_o)\ T^k(n_o)$$

where

$$C(n_o) = \frac{1}{\sum_{n=0}^{N-1} w(n)\, w(n+n_o)}$$

N is the number of samples in the window and typically has a value of 101, and $C(n_o)$ compensates for the window roll-off as a function of increasing autocorrelation lag. For non-integer values of n_o, the voiced energy at the nearest three values of n are used with a parabolic interpolation method to obtain the voiced energy for n_o. The total energy is determined as the voiced energy for n_o equal to zero.

With reference to Fig. 7, when speech signal s(n) enters a channel processing unit 44, components $s^i(n)$ belonging to a particular frequency band are isolated by a bandpass filter 48. Bandpass filter 48 uses downsampling to reduce computational requirements, and does so without any significant impact on system performance. Bandpass filter 48 can be implemented as a Finite Impulse Response (FIR) or Infinite Impulse Response (IIR) filter, or by using an FFT. A downsampling factor of S is achieved by shifting the input speech samples by S each time the filter outputs are computed.

A nonlinear operation unit 50 then performs a nonlinear operation on the isolated frequency band $s^i(n)$ to emphasize the fundamental frequency of the isolated frequency band $s^i(n)$. For complex values of $s^i(n)$ (i greater than zero), the absolute value, $|s^i(n)|$, is used. For the real value of $s^0(n)$, no nonlinear operation is performed.

The output of nonlinear operation unit 50 is passed through a highpass filter 52, and the output of the highpass filter is passed through an autocorrelation unit 54. A 101 point window is used, and, to reduce computation, the autocorrelation is only computed at a few samples nearest the pitch period.

With reference again to Fig. 4, second parameter estimator 16 may also use other approaches to produce the second voiced/unvoiced estimate. For example, well-known techniques such as using the height of the peak of the cepstrum, using the height of the peak of the autocorrelation of a linear prediction coder residual, MBE model parameter estimation methods, or IMBE (TM) model parameter estimation methods may be used. In addition, with reference again to Fig. 5, window and correlate unit 42 may produce autocorrelation values for the isolated frequency band $s^i(n)$ as:

$$R^i(l) = Re\left[\sum_n s^i(n+l)\, w(n+l)\, s^{*i}(n)\, w(n)\right]$$

where w(n) is the window. With this approach, combination block 18 produces the voiced/unvoiced parameters as:

$$V^k = \min(A^k, B^k).$$

The fundamental frequency may be estimated using a number of approaches. First, with reference to Fig. 8, a fundamental frequency estimation unit 56 includes a combining unit 58 and an estimator 60. Combining unit 58 sums the $T^i(\omega)$ outputs of channel processing units 20 (Fig. 2) to produce $X(\omega)$. In an alternative approach, combining unit 58 could estimate a signal-to-noise ratio (SNR) for the output of each channel processing unit 20 and weigh the various

outputs so that an output with a higher SNR contributes more to $X(\omega)$ than does an output with a lower SNR.

Estimator 60 then estimates the fundamental frequency (ω_o) by selecting a value for ω_o that maximizes $X(\omega_o)$ over an interval from ω_{min} to ω_{max}. Since $X(\omega)$ is only available at discrete samples of ω, parabolic interpolation of $X(\omega_o)$ near ω_o is used to improve accuracy of the estimate. Estimator 60 further improves the accuracy of the fundamental estimate by combining parabolic estimates near the peaks of the N harmonics of ω_o within the bandwidth of $X(\omega)$.

Once an estimate of the fundamental frequency is determined, the voiced energy $E^v(\omega_o)$ is computed as:

$$E^v(\omega_o) = \sum_{n=1}^{N} \sum_{\omega_m \in I_n} X(\omega_m)$$

where

$$I_n = [(n-0.25)\omega_o, (n+0.25)\omega_o].$$

Thereafter, the voiced energy $E^v(0.5\omega_o)$ is computed and compared to $E^v(\omega_o)$ to select between ω_o and $0.5\omega_o$ as the final estimate of the fundamental frequency.

With reference to Fig. 9, an alternative fundamental frequency estimation unit 62 includes a nonlinear operation unit 64, a windowing and Fast Fourier Transform (FFT) unit 66, and an estimator 68. Nonlinear operation unit 64 performs a nonlinear operation, the absolute value squared, on s(n) to emphasize the fundamental frequency of s(n) and to facilitate determination of the voiced energy when estimating ω_o.

Windowing and FFT unit 66 multiplies the output of nonlinear operation unit 64 to segment it and computes an FFT, $X(\omega)$, of the resulting product. Finally, estimator 68, which works identically to estimator 60, generates an estimate of the fundamental frequency.

With reference to Fig. 10, a hybrid fundamental frequency estimation unit 70 includes a band combination and estimation unit 72, an IMBE estimation unit 74 and an estimate combination unit 76. Band combination and estimation unit 70 combines the outputs of channel processing units 20 (Fig. 2) using simple summation or a signal-to-noise ratio (SNR) weighting where bands with higher SNRs are given higher weight in the combination. From the combined signal ($U(\omega)$), unit 72 estimates a fundamental frequency and a probability that the fundamental frequency is correct. Unit 72 estimates the fundamental frequency by choosing the frequency that maximizes the voiced energy ($E_v(\omega_o)$) from the combined signal, which is determined as:

$$E_v(\omega_o) = \sum_{n=1}^{N} \sum_{\omega_m \in I_n} U(\omega_m)$$

where

$$I_n = [(n-0.25)\omega_o, (n+0.25)\omega_o].$$

and N is the number of harmonics of the fundamental frequency. The probability that ω_o is correct is estimated by comparing $E_v(\omega_o)$ to the total energy E_t, which is computed as:

$$E_t = \sum_{\forall \omega_m > 0.5\omega_o} U(\omega_m) .$$

When $E_v(\omega_o)$ is close to E_t, the probability estimate is near one. When $E_v(\omega_o)$ is close to one half of E_t, the probability estimate is near zero.

IMBE estimation unit 74 uses the well known IMBE technique, or a similar technique, to produce a second fundamental frequency estimate and probability of correctness. Thereafter, estimate combination unit 76 combines the two

fundamental frequency estimates to produce the final fundamental frequency estimate. The probabilities of correctness are used so that the estimate with higher probability of correctness is selected or given the most weight.

With reference to Fig. 11, a voiced/unvoiced parameter smoothing unit 78 performs a smoothing operation to remove voicing errors that might result from rapid transitions in the speech signal. Unit 78 produces a smoothed voiced/unvoiced parameter as:

$$v_s^k(n) = 1.0, \text{ when } v^k(n\text{-}1)v^k(n\text{+}1) = 1 \text{ and}$$

$$v^k(n), \text{ otherwise}$$

where the voiced/unvoiced parameters equal zero for unvoiced speech and one for voiced speech. When the voiced/unvoiced parameters have continuous values, with a value near zero corresponding to highly voiced speech, unit 78 produces a smoothed voiced/unvoiced parameter that is smoothed in both the time and frequency domains:

$$v_s^k(n) = \lambda^k(n)\min(v^k(n), \alpha^k(n), \beta^k(n), \gamma^k(n))$$

where

$$\alpha^k(n) = 2v^{k+1}(n), \text{ when } k=0,1,\ldots,K\text{-}1, \text{ or}$$

$$\infty, \text{ when } k=K;$$

$$\beta^k(n) = 2v^{k-1}(n), \text{ when } k=2, 3,\ldots, K, \text{ or}$$

$$\infty, \text{ when } k=0,1;$$

$$\gamma^k(n) = 0.25v^{k-1}(n) + 0.5v^k(n) + 0.25v^{k+1}(n),$$

$$\text{when } k=1, 2, \ldots, K\text{-}1, \text{ or}$$

$$\infty, \text{ when } k=0, K;$$

$$\lambda^k(n) = 0.8, \text{ when } v_s^k(n\text{-}1) < T^k(n\text{-}1) \text{ and}$$

$$|\omega_o(n)\text{-}\omega_o(n\text{-}1)| < 0.25\,|\omega_o(n)|, \text{ or}$$

$$1, \text{ otherwise;}$$

and $T^k(n)$ is a threshold value that is a function of time and frequency.

With reference to Fig. 12, a voiced/unvoiced parameter improvement unit 80 produces improved voiced/unvoiced parameters by comparing the voiced/unvoiced parameter produced when the estimated fundamental frequency equals ω_o to a voiced/unvoiced parameter produced when the estimated fundamental frequency equals one half of ω_o and selecting the parameter having the lowest value. In particular, voiced/unvoiced parameter improvement unit 80 produces improved voiced/unvoiced parameters as:

$$A^k(\omega_o) = \min(A^k(\omega_o), A^k(0.5\omega_o))$$

where

$$A^k(\omega) = 1.0\text{-}\ E_v^k(\omega)/E_t^k.$$

With reference to Fig. 13, an improved estimate of the fundamental frequency (ω_o) is generated according to a procedure 100. The initial fundamental frequency estimate ($\tilde{\omega}_o$) is generated according to one of the procedures described above and is used in step 101 to generate a set of evaluation frequencies $\tilde{\omega}^k$. The evaluation frequencies are typically chosen to be near the integer submultiples and multiples of $\tilde{\omega}_o$. Thereafter, functions are evaluated at this set of evaluation frequencies (step 102). The functions that are evaluated typically consist of the voiced energy function $E_v(\tilde{\omega}^k)$ and the normalized frame error $E_f(\tilde{\omega}^k)$. The normalized frame error is computed as

$$E_f(\omega^k) = 1.0 - E_v(\omega^k) / E_t(\omega^k) .$$

The final fundamental frequency estimate is then selected (step 103) using the evaluation frequencies, the function values at the evaluation frequencies, the predicted fundamental frequency (described below), the final fundamental frequency estimates from previous frames, and the above function values from previous frames. When these inputs indicate that one evaluation frequency has a much higher probability of being the correct fundamental frequency than the others, then it is chosen. Otherwise, if two evaluation frequencies have similar probability of being correct and the normalized error for the previous frame is relatively low, then the evaluation frequency closest to the final fundamental frequency from the previous frame is chosen. Otherwise, if two evaluation frequencies have similar probability of being correct, then the one closest to the predicted fundamental frequency is chosen. The predicted fundamental frequency for the next frame is generated (step 104) using the final fundamental frequency estimates from the current and previous frames, a delta fundamental frequency, and normalized frame errors computed at the final fundamental frequency estimate for the current frame and previous frames. The delta fundamental frequency is computed from the frame to frame difference in the final fundamental frequency estimate when the normalized frame errors for these frames are relatively low and the percentage change in fundamental frequency is low, otherwise, it is computed from previous values. When the normalized error for the current frame is relatively low, the predicted fundamental for the current frame is set to the final fundamental frequency. The predicted fundamental for the next frame is set to the sum of the predicted fundamental for the current frame and the delta fundamental frequency for the current frame.

Claims

1. A method of analysing a digitized speech signal to determine excitation parameters for the digitized speech signal, preferably as a step in encoding speech, the method comprising dividing the digitized speech signal into one or more frequency band signals; and, preferably at regular intervals of time, performing the further step of: determining a first preliminary excitation parameter using a first method that includes performing a nonlinear operation on at least one of the frequency band signals to produce at least one modified frequency band signal and determining the first preliminary excitation parameter using the at least one modified frequency band signal; determining at least a second preliminary excitation parameter using at least a second method different from the said first method; and using the first and at least a second preliminary excitation parameters to determine an excitation parameter for the digitized speech signal.

2. A method according to Claim 1, wherein at least one of the second methods uses at least one of the frequency band signals without performing the said nonlinear operation.

3. A method according to Claims 1 or 2, wherein the excitation parameter comprises a voiced/unvoiced parameter for at least one frequency band, said parameter preferably having values that vary over a continuous range.

4. A method according to any preceding claim, further comprising determining a fundamental frequency for the digitized speech signal.

5. A method according to Claim 3, wherein the first preliminary excitation parameter comprises a first voiced/unvoiced parameter for the at least one modified frequency band signal, and wherein the first determining step includes determining the first voiced/unvoiced parameter by comparing voiced energy in the modified frequency band signal to total energy in the modified frequency band signal.

6. A method according to Claim 5, wherein the voiced energy in the modified frequency band signal corresponds to the energy associated with an estimated fundamental frequency for the digitized speech signal.

7. A method according to Claim 5, wherein the voiced energy in the modified frequency band signal corresponds to the energy associated with an estimated pitch period for the digitized speech signal.

8. A method according to Claim 5, wherein the second preliminary excitation parameter includes a second voiced/unvoiced parameter for the at least one frequency band signal, and wherein the second determining step includes determining the second voiced/unvoiced parameter by comparing sinusoidal energy in the at least one frequency band signal to total energy in the at least one frequency band signal.

9. A method according to Claim 5, wherein the second preliminary excitation parameter includes a second voiced/unvoiced parameter for the at least one frequency band signal, and wherein the second determining step includes determining the second voiced/unvoiced parameter by autocorrelating the at least one frequency band signal.

10. A method according to any preceding claim, wherein the said using step emphasizes the first preliminary excitation parameter over the second preliminary excitation parameter in determining the excitation parameter for the digitized speech signal when the first preliminary excitation parameter has a higher probability of being correct than does the second preliminary excitation parameter.

11. A method according to any preceding claim, further comprising smoothing the excitation parameter to produce a smoothed excitation parameter.

12. A method of analysing a digitized speech signal to determine excitation parameters for the digitized speech signal, preferably as a step in encoding speech, the method comprising the steps of: determining preliminary excitation parameters from the digitized speech signal; and smoothing the preliminary excitation parameters to produce excitation parameters.

13. A method according to Claim 12, wherein the preliminary excitation parameters include a preliminary voiced/unvoiced parameter for at least one frequency band and the excitation parameters include a voiced/unvoiced parameter for at least one frequency band, which voiced/unvoiced parameter preferably has values that vary over a continuous range.

14. A method according to Claim 13, wherein the excitation parameters include a fundamental frequency.

15. A method according to Claims 13 or 14, wherein the smoothing step makes the voiced/unvoiced parameter more voiced than the preliminary voiced/unvoiced parameter when voiced/unvoiced parameters that are nearby in time and/or frequency are voiced.

16. A method according to Claim 12, wherein the smoothing step is performed as a function of time and/or frequency.

17. A method of analysing a digitized speech signal to determine excitation parameters for the digitized speech signal, preferably as a step in encoding speech, the method comprising the steps of: estimating a fundamental frequency for the digitized speech signal; evaluating a voiced/unvoiced function using the estimated fundamental frequency to produce a first preliminary voiced/unvoiced parameter; evaluating the voiced/unvoiced function at least using one other frequency derived from the estimated fundamental frequency to produce at least one other preliminary voiced/unvoiced parameter; and combining the first and at least one other preliminary voiced/unvoiced parameters to produce a voiced/unvoiced parameter.

18. A method according to Claim 17, wherein the said at least one other frequency is derived from the said estimated fundamental frequency as a multiple or submultiple of the said estimated fundamental frequency.

19. A method according to Claim 17, wherein the combining step includes choosing the first preliminary voiced/unvoiced parameter as the voiced/unvoiced parameter when the first preliminary voiced/unvoiced parameter indicates that the digitized speech signal is more voiced than does the second preliminary voiced/unvoiced parameter.

20. A method of synthesizing speech using excitation parameters, where the excitation parameters are estimated by using a method for determining such parameters according to any preceding claim.

21. A method of analysing a digitized speech signal to determine a fundamental frequency estimate for the digitized speech signal, comprising the steps of: determining a predicted fundamental frequency estimate from previous

fundamental frequency estimates; determining an initial fundamental frequency estimate; evaluating an error function at the initial fundamental frequency estimate to produce a first error function value; evaluating the error function at at least one other frequency derived from the initial fundamental frequency estimate to produce at least one other error function value; selecting a fundamental frequency estimate using the predicted fundamental frequency estimate, the initial fundamental frequency estimate, the first error function value, and the at least one other error function value.

22. A method according to Claim 21, wherein the said at least one other frequency is derived from the said estimated fundamental frequency as a multiple or submultiple of the said estimated fundamental frequency.

23. A method according to Claim 21, wherein the predicted fundamental frequency is determined by adding a delta factor to a previous predicted fundamental frequency, which delta factor is preferably determined from previous first and at least one other error function values, the previous predicted fundamental frequency, and a previous delta factor.

24. A method of synthesizing speech using a fundamental frequency, where the fundamental frequency is estimated using a method according to any of Claims 21, 22 or 23.

25. A system for analysing a digitized speech signal to determine excitation parameters for the digitized speech signal, comprising: means for dividing the digitized speech signal into one or more frequency band signals; means for determining a first preliminary excitation parameter using a first method that includes performing a nonlinear operation on at least one of the frequency band signals to produce at least one modified frequency band signal and determining the first preliminary excitation parameter using the at least one modified frequency band signal; means for determining a second preliminary excitation parameter using a second method that is different from the above said first method; and means for using the first and second preliminary excitation parameters to determine an excitation parameter for the digitized speech signal.

26. A system for analysing a digitized speech signal to determine excitation parameters for the digitized speech signal, comprising: means for determining preliminary excitation parameters from the digitized speech signal; and means for smoothing the preliminary excitation parameters to produce excitation parameters.

27. A system for analysing a digitized speech signal to determine modified excitation parameters for the digitized speech signal, comprising: means for estimating a fundamental frequency for the digitized speech signal; means for evaluating a voiced/unvoiced function using the estimated fundamental frequency to produce a first preliminary voiced/unvoiced parameter; means for evaluating the voiced/unvoiced function using another frequency derived from the estimated fundamental frequency to produce a second preliminary voiced/unvoiced parameter; and means for combining the first and second preliminary voiced/unvoiced parameters to produce a voiced/unvoiced parameter.

28. A system for analysing a digitized speech signal to determine a fundamental frequency estimate for the digitized speech signal, comprising: means for determining a predicted fundamental frequency estimate from previous fundamental frequency estimates; means for determining an initial fundamental frequency estimate; means for evaluating an error function at the initial fundamental frequency estimate to produce a first error function value; means for evaluating the error function at at least one other frequency derived from the initial fundamental frequency estimate to produce a second error function value; and means for selecting a fundamental frequency estimate using the predicted fundamental frequency estimate, the initial fundamental frequency estimate, the first error function value, and the second error function value.

29. A method of analysing a digitized speech signal to determine a voiced/unvoiced function for the digitized speech signal, comprising: dividing the digitized speech signal into at least two frequency band signals; determining a first preliminary voiced/unvoiced function for at least two of the frequency band signals using a first method; determining a second preliminary voiced/unvoiced function for at least two of the frequency band signals using a second method which is different from the above said first method; and using the first and second preliminary excitation parameters to determine a voiced/unvoiced function for at least two of the frequency band signals.

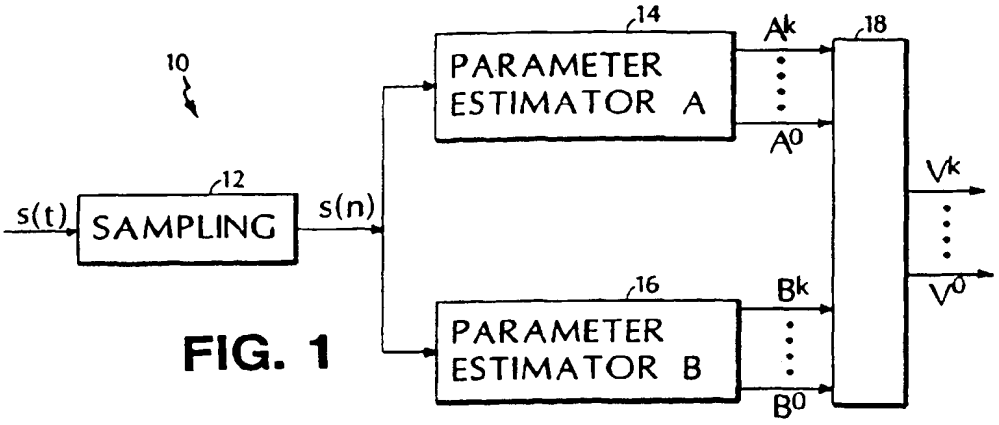

FIG. 1

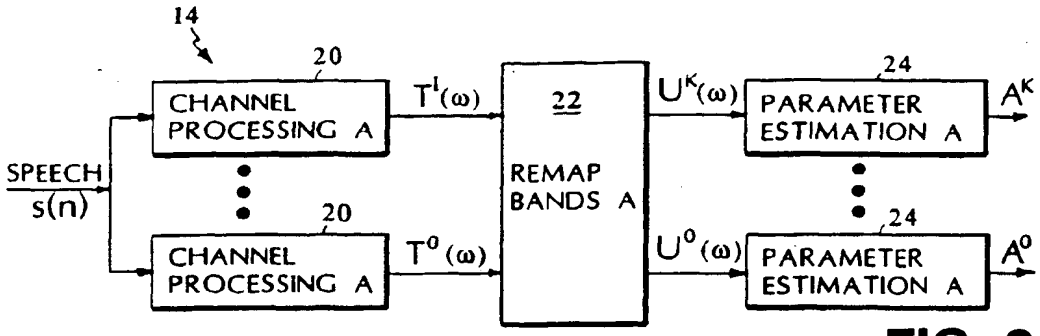

FIG. 2

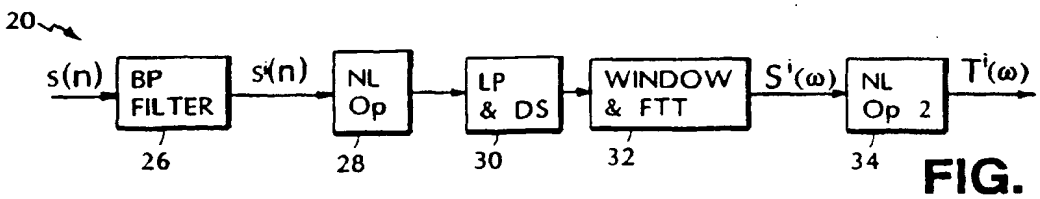

FIG. 3

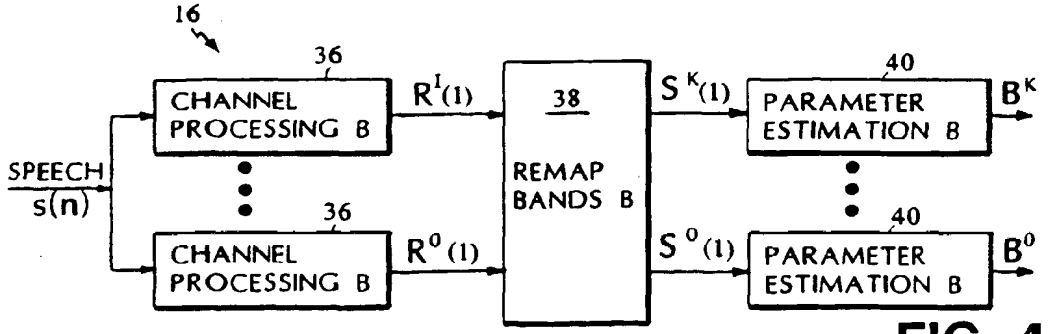

FIG. 4

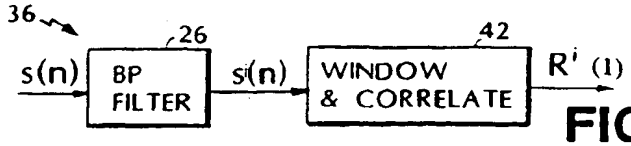

FIG. 5

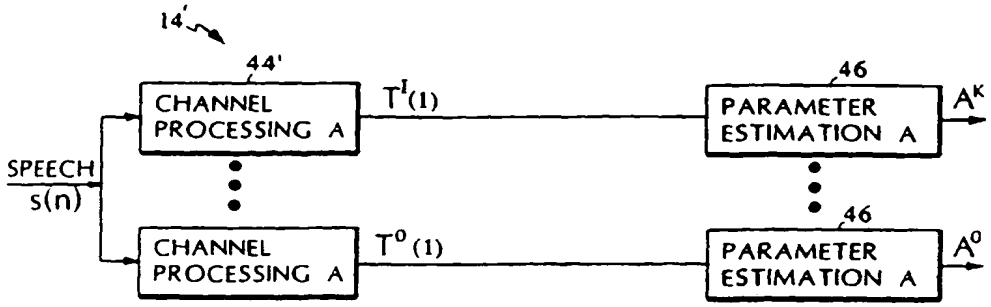

FIG. 6

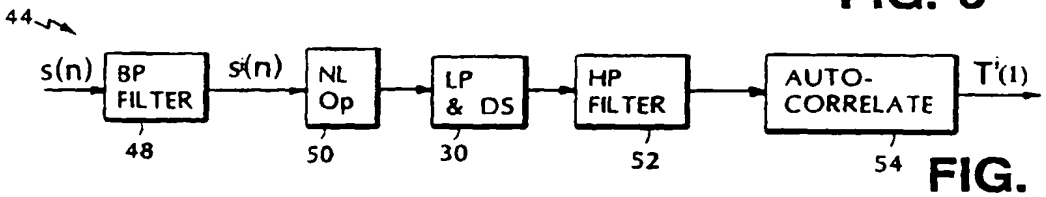

FIG. 7

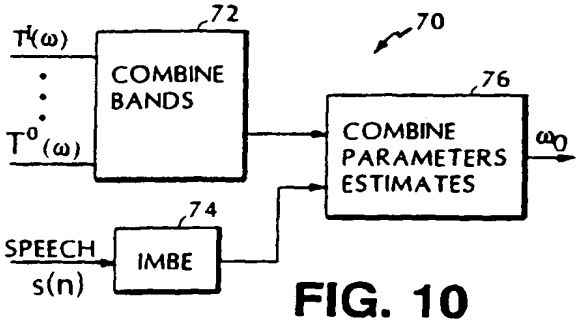

FIG. 10

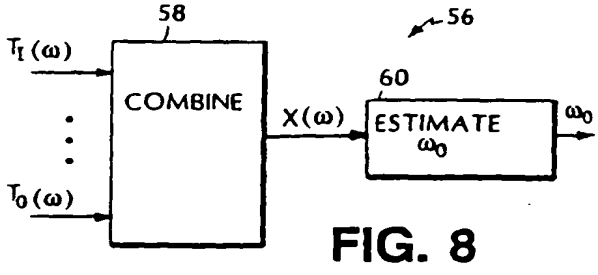

FIG. 8

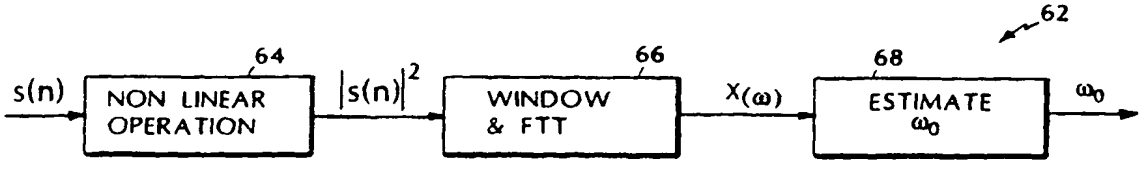

FIG. 9

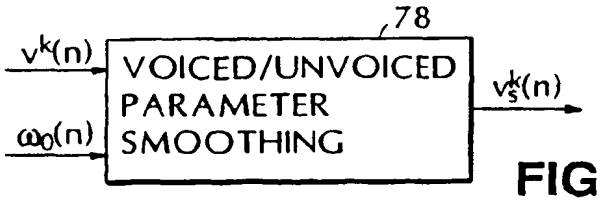

FIG. 11

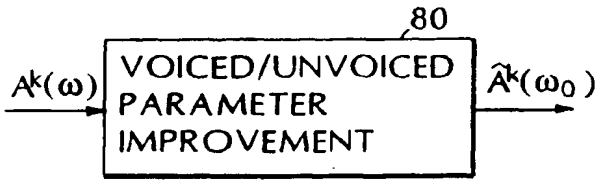

FIG. 12

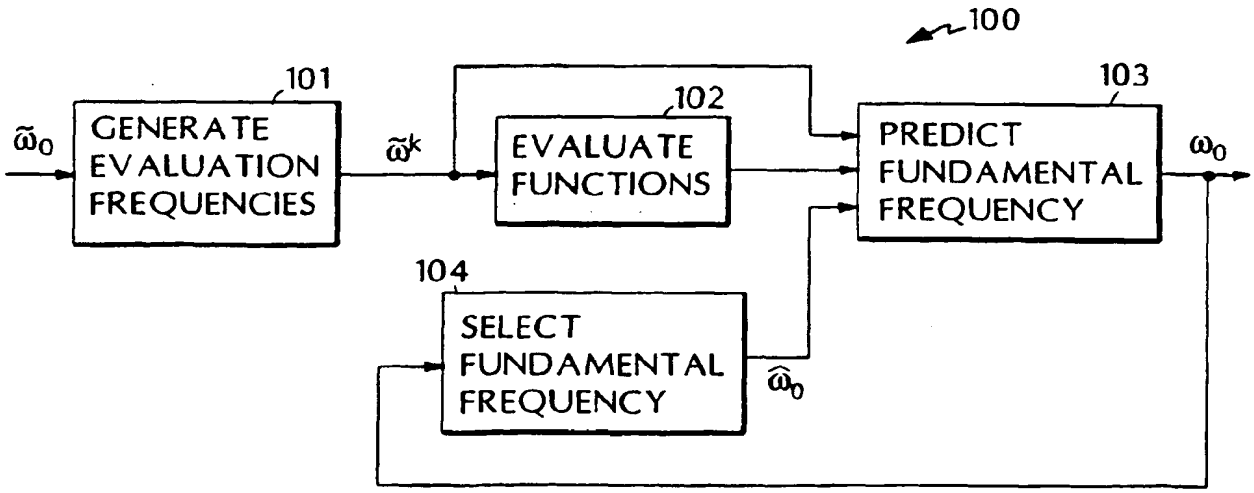

FIG. 13

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 722 165 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
17.07.1996 Bulletin 1996/29

(51) Int Cl.[6]: **G10L 9/14**

(21) Application number: **96300245.6**

(22) Date of filing: **12.01.1996**

(84) Designated Contracting States:
DE FR GB SE

(30) Priority: **12.01.1995 US 371743**

(71) Applicant: **DIGITAL VOICE SYSTEMS, INC.**
Burlington, MA 01803 (US)

(72) Inventor: **Griffin, Daniel Wayne**
Hollis, New Hampshire 03049 (US)

(74) Representative: **Deans, Michael John Percy et al**
Lloyd Wise, Tregear & Co.,
Commonwealth House,
1-19 New Oxford Street
London WC1A 1LW (GB)

(54) Estimation of excitation parameters

(57) Excitation parameters for a digitized speech signal are determined by analysing the digitized speech signal. The digitized speech signal is divided into at least two frequency bands. A first preliminary excitation parameter is determined by performing a nonlinear operation on at least one of the frequency band signals to produce a modified frequency band signal and determining the first preliminary excitation parameter using the modified frequency band signal. A second preliminary excitation parameter is determined using a method different from the first method. The first and second preliminary excitation parameters are used to determine an excitation parameter for the digitized speech signal. The method is useful in encoding speech. Speech synthesized using the parameters estimated based on the invention generates high quality speech at various bit rates useful for applications such as satellite voice communication.

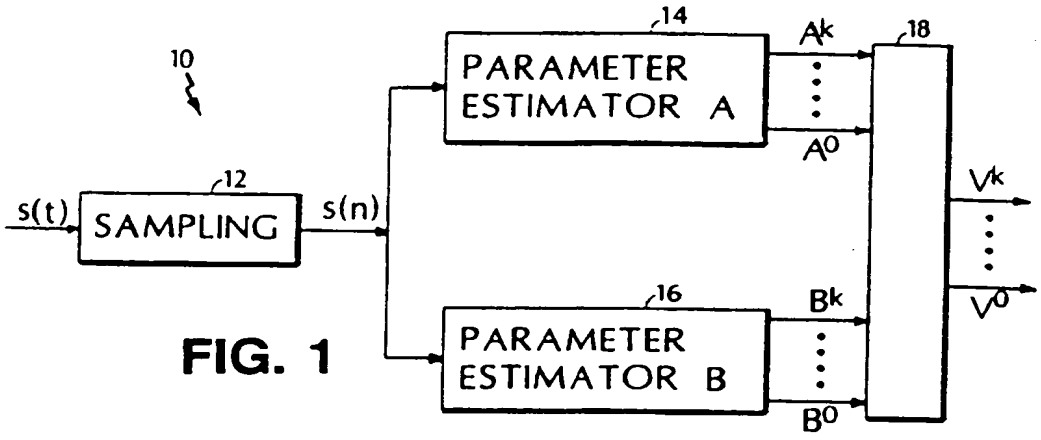

FIG. 1

EP 0 722 165 A2 CORRIGENDUM issued on 20.11.1996 (bibliography updates included)

Description

The invention has arisen from work seeking to improve the accuracy with which excitation parameters are estimated in speech analysis and synthesis.

Speech analysis and synthesis are widely used in applications such as telecommunications and voice recognition. A vocoder, which is a type of speech analysis/synthesis system, models speech as the response of a system to excitation over short time intervals. Examples of vocoder systems include linear prediction vocoders, homomorphic vocoders, channel vocoders, sinusoidal transform coders ("STC"), multiband excitation ("MBE") vocoders, improved multiband excitation ("IMBE (TM)") vocoders.

Vocoders typically synthesize speech based on excitation parameters and system parameters. Typically, an input signal is segmented using, for example, a Hamming window. Then, for each segment, system parameters and excitation parameters are determined. System parameters include the spectral envelope or the impulse response of the system. Excitation parameters include a fundamental frequency (or pitch) and a voiced/unvoiced parameter that indicates whether the input signal has pitch (or indicates the degree to which the input signal has pitch). In vocoders that divide the speech into frequency bands, such as IMBE (TM) vocoders, the excitation parameters may also include a voiced/unvoiced parameter for each frequency band rather than a single voiced/unvoiced parameter. Accurate excitation parameters are essential for high quality speech synthesis.

When the voiced/unvoiced parameters include only a single voiced/unvoiced decision for the entire frequency band, the synthesized speech tends to have a "buzzy" quality especially noticeable in regions of speech which contain mixed voicing or in voiced regions of noisy speech. A number of mixed excitation models have been proposed as potential solutions to the problem of "buzziness" in vocoders. In these models, periodic and noise-like excitations are mixed which have either time-invariant or time-varying spectral shapes.

In excitation models having time-invariant spectral shapes, the excitation signal consists of the sum of a periodic source and a noise source with fixed spectral envelopes. The mixture ratio controls the relative amplitudes of the periodic and noise sources. Examples of such models include Itakura and Saito, "Analysis Synthesis Telephony Based upon the Maximum Likelihood Method," Reports of 6th Int. Cong. Acoust., Tokyo, Japan, Paper C-5-5, pp. C17-20, 1968; and Kwon and Goldberg, "An Enhanced LPC Vocoder with No Voiced/Unvoiced Switch," IEEE Trans. on Acoust., Speech, and Signal Processing, vol. ASSP-32, no. 4, pp. 851-858, August 1984. In theses excitation models a white noise source is added to a white periodic source. The mixture ratio between these sources is estimated from the height of the peak of the autocorrelation of the LPC residual.

In excitation models having time-varying spectral shapes, the excitation signal consists of the sum of a periodic source and a noise source with time varying spectral envelope shapes. Examples of such models include Fujimara, "An Approximation to Voice Aperiodicity," IEEE Trans. Audio and Electroacoust., pp. 68-72, March 1968; Makhoul et al., "A Mixed-Source Excitation Model for Speech Compression and Synthesis," IEEE Int. Conf. on Acoust. Sp. & Sig. Proc., April 1978, pp. 163-166; Kwon and Goldberg, "An Enhanced LPC Vocoder with No Voiced/Unvoiced Switch," IEEE Trans. on Acoust., Speech, and Signal Processing, vol. ASSP-32, no. 4, pp. 851-858, August 1984; and Griffin and Lim, "Multiband Excitation Vocoder," IEEE Trans.
Acoust., Speech, Signal Processing, vol. ASSP-36, pp. 1223-1235, Aug. 1988.

In the excitation model proposed by Fujimara, the excitation spectrum is divided into three fixed frequency bands A separate cepstral analysis is performed for each frequency band and a voiced/unvoiced decision for each frequency band is made based on the height of the cepstrum peak as a measure of periodicity.

In the excitation model proposed by Makhoul et al., the excitation signal consists of the sum of a low-pass periodic source and a high-pass noise source. The low-pass periodic source is generated by filtering a white pulse source with a variable cut-off low-pass filter. Similarly, the high-pass noise source was generated by filtering a white noise source with a variable cut-off high-pass filter. The cut-off frequencies for the two filters are equal and are estimated by choosing the highest frequency at which the spectrum is periodic. Periodicity of the spectrum is determined by examining the separation between consecutive peaks and determining whether the separations are the same, within some tolerance level.

In a second excitation model implemented by Kwon and Goldberg, a pulse source is passed through a variable gain low-pass filter and added to itself, and a white noise source is passed through a variable gain high-pass filter and added to itself. The excitation signal is the sum of the resultant pulse and noise sources with the relative amplitudes controlled by a voiced/unvoiced mixture ratio. The filter gains and voiced/unvoiced mixture ratio are estimated from the LPC residual signal with the constraint that the spectral envelope of the resultant excitation signal is flat.

In the multiband excitation model proposed by Griffin and Lim, a frequency dependent voiced/unvoiced mixture function is proposed. This model is restricted to a frequency dependent binary voiced/unvoiced decision for coding purposes. A further restriction of this model divides the spectrum into a finite number of frequency bands with a binary voiced/unvoiced decision for each band. The voiced/unvoiced information is estimated by comparing the speech spectrum to the closest periodic spectrum. When the error is below a threshold, the band is marked voiced, otherwise, the

band is marked unvoiced.

Excitation parameters may also be used in applications, such as speech recognition, where no speech synthesis is required. Once again, the accuracy of the excitation parameters directly affects the performance of such a system.

In one aspect, generally, the invention features a hybrid excitation parameter estimation technique that produces two sets of excitation parameters for a speech signal using two different approaches and combines the two sets to produce a single set of excitation parameters. In a first approach, the technique applies a nonlinear operation to the speech signal to emphasize the fundamental frequency of the speech signal. In a second approach, we use a different method which may or may not include a nonlinear operation. While the first approach produces highly accurate excitation parameters under most conditions, the second approach produces more accurate parameters under certain conditions. By using both approaches and combining the resulting sets of excitation parameters to produce a single set, our technique produces accurate results under a wider range of conditions than are produced by either of the approaches individually.

In typical approaches to determining excitation parameters, an analog speech signal s(t) is sampled to produce a speech signal s(n). Speech signal s(n) is then multiplied by a window w(n) to produce a windowed signal $s_w(n)$ that is commonly referred to as a speech segment or a speech frame. A Fourier transform is then performed on windowed signal $s_w(n)$ to produce a frequency spectrum $S_w(\omega)$ from which the excitation parameters are determined.

When speech signal s(n) is periodic with a fundamental frequency ω_o or pitch period n_o (where n_o equals $2\pi/\omega_o$), the frequency spectrum of speech signal s(n) should be a line spectrum with energy at ω_o and harmonics thereof (integral multiples of ω_o). As expected, $S_w(\omega)$ has spectral peaks that are centered around ω_o and its harmonics. However, due to the windowing operation, the spectral peaks include some width, where the width depends on the length and shape of window w(n) and tends to decrease as the length of window w(n) increases. This window-induced error reduces the accuracy of the excitation parameters. Thus, to decrease the width of the spectral peaks, and to thereby increase the accuracy of the excitation parameters, the length of window w(n) should be made as long as possible.

The maximum useful length of window w(n) is limited. Speech signals are not stationary signals, and instead have fundamental frequencies that change over time. To obtain meaningful excitation parameters, an analyzed speech segment must have a substantially unchanged fundamental frequency. Thus, the length of window w(n) must be short enough to ensure that the fundamental frequency will not change significantly within the window.

In addition to limiting the maximum length of window w(n), a changing fundamental frequency tends to broaden the spectral peaks. This broadening effect increases with increasing frequency. For example, if the fundamental frequency changes by $\Delta\omega_o$ during the window, the frequency of the *mth* harmonic, which has a frequency of $m\omega_o$, changes by $m\Delta\omega_o$ so that the spectral peak corresponding to $m\omega_o$ is broadened more than the spectral peak corresponding to ω_o. This increased broadening of the higher harmonics reduces the effectiveness of higher harmonics in the estimation of the fundamental frequency and the generation of voiced/unvoiced parameters for high frequency bands.

By applying a nonlinear operation to the speech signal, the increased impact on higher harmonics of a changing fundamental frequency is reduced or eliminated, and higher harmonics perform better in estimation of the fundamental frequency and determination of voiced/unvoiced parameters. Suitable nonlinear operations map from complex (or real) to real values and produce outputs that are nondecreasing functions of the magnitudes of the complex (or real) values. Such operations include, for example, the absolute value, the absolute value squared, the absolute value raised to some other power, or the log of the absolute value.

Nonlinear operations tend to produce output signals having spectral peaks at the fundamental frequencies of their input signals. This is true even when an input signal does not have a spectral peak at the fundamental frequency. For example, if a bandpass filter that only passes frequencies in the range between the third and fifth harmonics of ω_o is applied to a speech signal s(n), the output of the bandpass filter, x(n), will have spectral peaks at $3\omega_o$, $4\omega_o$ and $5\omega_o$.

Though x(n) does not have a spectral peak at ω_o, $|x(n)|^2$ will have such a peak. For a real signal x(n), $|x(n)|^2$ is equivalent to $x^2(n)$. As is well known, the Fourier transform of $x^2(n)$ is the convolution of X(ω), the Fourier transform of x(n), with X(ω):

$$\sum_{n=-\infty}^{\infty} x^2(n)\, e^{-j\omega n} = \frac{1}{2\pi} \int_{u=-\pi}^{\pi} X(\omega-u)\, X(u)\, du.$$

The convolution of X(ω) with X(ω) has spectral peaks at frequencies equal to the differences between the frequencies for which X(ω) has spectral peaks. The differences between the spectral peaks of a periodic signal are the fundamental frequency and its multiples. Thus, in the example in which X(ω) has spectral peaks at $3\omega_o$, $4\omega_o$ and $5\omega_o$, X(ω) convolved with X(ω) has a spectral peak at ω_o ($4\omega_o$-$3\omega_o$, $5\omega_o$-$4\omega_o$). For a typical periodic signal, the spectral peak at the funda-

mental frequency is likely to be the most prominent.

The above discussion also applies to complex signals. For a complex signal x(n), the Fourier transform of $|x(n)|^2$ is:

$$\sum_{n=-\infty}^{\infty} |x(n)|^2 e^{-j\omega n} = \frac{1}{2\pi} \int_{u=-\pi}^{\pi} X(\omega+u)\, X^*(u)\, du.$$

This is an autocorrelation of $X(\omega)$ with $X^*(\omega)$, and also has the property that spectral peaks separated by $n\omega_o$ produce peaks at $n\omega_o$.

Even though $|x(n)|$, $|x(n)|^a$ for some real "a", and $\log |x(n)|$ are not the same as $|x(n)|^2$, the discussion above for $|x(n)|^2$ applies approximately at the qualitative level. For example, for $|x(n)| = y(n)^{0.5}$, where $y(n) = |x(n)|^2$, a Taylor series expansion of y(n) can be expressed as:

$$|x(n)| = \sum_{k=0}^{\infty} c_k y^k(n) .$$

Because multiplication is associative, the Fourier transform of the signal $y^k(n)$ is $Y(\omega)$ convolved with the Fourier transform of $y^{k-1}(n)$. The behavior for nonlinear operations other than $|x(n)|^2$ can be derived from $|x(n)|^2$ by observing the behavior of multiple convolutions of $Y(\omega)$ with itself. If $Y(\omega)$ has peaks at $n\omega_o$, then multiple convolutions of $Y(\omega)$ with itself will also have peaks at $n\omega_o$.

As shown, nonlinear operations emphasize the fundamental frequency of a periodic signal, and are particularly useful when the periodic signal includes significant energy at higher harmonics. However, the presence of the nonlinearity can degrade performance in some cases. For example, performance may be degraded when speech signal s (n) is divided into multiple bands $s^i(n)$ using bandpass filters, where $s^i(n)$ denotes the result of bandpass filtering using the ith bandpass filter. If a single harmonic of the fundamental frequency is present in the pass band of the ith filter, the output of the filter is:

$$s^i(n) = A_k e^{j(\omega_k n + \theta_k)}$$

where ω_k is the frequency, θ_k is the phase, and A_k is the amplitude of the harmonic. When a nonlinearity such as the absolute value is applied to $s^i(n)$ to produce a value $y^i(n)$, the result is:

$$y^i(n) = |s^i(n)| = |A_k|$$

so that the frequency information has been completely removed from the signal $y^i(n)$. Removal of this frequency information can reduce the accuracy of parameter estimates.

Our hybrid technique provides significantly improved parameter estimation performance in cases for which the nonlinearity reduces the accuracy of parameter estimates while maintaining the benefits of the nonlinearity in the remaining cases. As described above, the hybrid technique includes combining parameter estimates based on the signal after the nonlinearity has been applied ($y^i(n)$) with parameter estimates based on the signal before the nonlinearity is applied ($s^i(n)$ or $s(n)$). The two approaches produce parameter estimates along with an indication of the probability of correctness of these parameter estimates. The parameter estimates are then combined giving higher weight to estimates with a higher probability of being correct.

In another aspect, generally, the invention features the application of smoothing techniques to the voiced/unvoiced parameters. Voiced/unvoiced parameters can be binary or continuous functions of time and/or frequency. Because these parameters tend to be smooth functions in at least one direction (positive or negative) of time or frequency, the estimates of these parameters can benefit from appropriate application of smoothing techniques in time and/or frequency.

The invention also features an improved technique for estimating voiced/unvoiced parameters. In vocoders such as linear prediction vocoders, homomorphic vocoders, channel vocoders, sinusoidal transform coders, multiband excitation vocoders, and IMBE (TM) vocoders, a pitch period n (or equivalently a fundamental frequency) is selected.

Thereafter, a function $f^i(n)$ is then evaluated at the selected pitch period (or fundamental frequency) to estimate the ith voiced/unvoiced parameter. However, for some speech signals, evaluation of this function only at the selected pitch period will result in reduced accuracy of one or more voiced/unvoiced parameter estimates. This reduced accuracy may result from speech signals that are more periodic at a multiple of the pitch period than at the pitch period, and may be frequency dependent so that only certain portions of the spectrum are more periodic at a multiple of the pitch period. Consequently, the voiced/unvoiced parameter estimation accuracy can be improved by evaluating the function $f^i(n)$ at the pitch period n and at its multiples, and thereafter combining the results of these evaluations.

In another aspect, the invention features an improved technique for estimating the fundamental frequency or pitch period. When the fundamental frequency ω_o (or pitch period n_o) is estimated, there may be some ambiguity as to whether ω_o or a submultiple or multiple of ω_o is the best choice for the fundamental frequency. Since the fundamental frequency tends to be a smooth function of time for voiced speech, predictions of the fundamental frequency based on past estimates can be used to resolve ambiguities and improve the fundamental frequency estimate.

Other features and advantages of the invention will be apparent from the following description of the preferred embodiments, given by way of example only.

In the drawings:

Fig. 1 is a block diagram of a system for determining whether frequency bands of a signal are voiced or unvoiced.

Fig. 2 is a block diagram of a parameter estimation unit of the system of Fig. 1.

Fig. 3 is a block diagram of a channel processing unit of the parameter estimation unit of Fig. 2.

Fig. 4 is a block diagram of a parameter estimation unit of the system of Fig. 1.

Fig. 5 is a block diagram of a channel processing unit of the parameter estimation unit of Fig. 4.

Fig. 6 is a block diagram of a parameter estimation unit of the system of Fig. 1.

Fig. 7 is a block diagram of a channel processing unit of the parameter estimation unit of Fig. 6.

Figs. 8-10 are block diagrams of systems for determining the fundamental frequency of a signal.

Fig. 11 is a block diagram of voiced/unvoiced parameter smoothing unit.

Fig. 12 is a block diagram of voiced/unvoiced parameter improvement unit.

Fig. 13 is a block diagram of a fundamental frequency improvement unit.

Figs. 1-12 show the structure of a system for estimating excitation parameters, the various blocks and units of which are preferably implemented with software.

With reference to Fig. 1, a voiced/unvoiced determination system 10 includes a sampling unit 12 that samples an analog speech signal s(t) to produce a speech signal s(n). For typical speech coding applications, the sampling rate ranges between six kilohertz and ten kilohertz.

Speech signal s(n) is supplied to a first parameter estimator 14 that divides the speech signal into k+1 bands and produces a first set of preliminary voiced/unvoiced ("V/UV") parameters (A^0 to A^K) corresponding to a first estimate as to whether the signals in the bands are voiced or unvoiced. Speech signal s(n) is also supplied to a second parameter estimator 16 that produces a second set of preliminary V/UV parameters (B^0 to B^K) that correspond to a second estimate as to whether the signals in the bands are voiced or unvoiced. The two sets of preliminary V/UV parameters are combined by a combination block 18 to produce a set of V/UV parameters (V^0 to V^K).

With reference to Fig. 2, first parameter estimator 14 produces the first voiced/unvoiced estimate using a frequency domain approach. Channel processing units 20 in first parameter estimator 14 divide speech signal s(n) into at least two frequency bands and process the frequency bands to produce a first set of frequency band signals, designated as $T^0(\omega)$ .. $T^I(\omega)$. As discussed below, channel processing units 20 are differentiated by the parameters of a bandpass filter used in the first stage of each channel processing unit 20. In the described embodiment, there are sixteen channel processing units (I equals 15).

A remap unit 22 transforms the first set of frequency band signals to produce a second set of frequency band signals, designated as $U^0(\omega)$ .. $U^K(\omega)$. In the described embodiment, there are eight frequency band signals in the second set of frequency band signals (K equals 7). Thus, remap unit 22 maps the frequency band signals from the sixteen channel processing units 20 into eight frequency band signals. Remap unit 20 does so by combining consecutive pairs of frequency band signals from the first set into single frequency band signals in the second set. For example, $T^0(\omega)$ and $T^1(\omega)$ are combined to produce $U^0(\omega)$, and $T^{14}(\omega)$ and $T^{15}(\omega)$ are combined to produce $U^7(\omega)$. Other approaches to remapping could also be used.

Next, voiced/unvoiced parameter estimation units 24, each associated with a frequency band signal from the second set, produce preliminary V/UV parameters A^0 to A^K by computing a ratio of the voiced energy in the frequency band at an estimated fundamental frequency ω^o to the total energy in the frequency band and subtracting this ratio from 1:

$$A^k = 1.0 - E_v^k(\omega_o) / E_t^k .$$

The voiced energy in the frequency band is computed as:

$$E_V^k(\omega_o) = \sum_{n=1}^{N} \sum_{\omega_m \in I_n} U^k(\omega_m)$$

where

$$I_n = [(n\text{-}0.25)\ \omega_o, (n\text{+}0.25)\ \omega_o],$$

and N is the number of harmonics of the fundamental frequency ω_o being considered. V/UV parameter estimation units 24 determine the total energy of their associated frequency band signals as:

$$E_T^k = \sum_{\forall \omega_m > 0.5\omega_o} U^k(\omega_m) .$$

The degree to which the frequency band signal is voiced varies indirectly with the value of the preliminary V/UV parameter. Thus, the frequency band signal is highly voiced when the preliminary V/UV parameter is near zero and is highly unvoiced when the parameter is greater than or equal to one half.

With reference to Fig. 3, when speech signal s(n) enters a channel processing unit 20, components $s^i(n)$ belonging to a particular frequency band are isolated by a bandpass filter 26. Bandpass filter 26 uses downsampling to reduce computational requirements, and does so without any significant impact on system performance. Bandpass filter 26 can be implemented as a Finite Impulse Response (FIR) or Infinite Impulse Response (IIR) filter, or by using an FFT. In the described embodiment, bandpass filter 26 is implemented using a thirty two point real input FFT to compute the outputs of a thirty two point FIR filter at seventeen frequencies, and achieves a downsampling factor of S by shifting the input by S samples each time the FFT is computed. For example, if a first FFT used samples one through thirty two, a downsampling factor of ten would be achieved by using samples eleven through forty two in a second FFT.

A first nonlinear operation unit 28 then performs a nonlinear operation on the isolated frequency band $s^i(n)$ to emphasize the fundamental frequency of the isolated frequency band $s^i(n)$. For complex values of $s^i(n)$ (i greater than zero), the absolute value, $|s^i(n)|$, is used. For the real value of $s^0(n)$, $s^0(n)$ is used if $s^0(n)$ is greater than zero and zero is used if $s^0(n)$ is less than or equal to zero.

The output of nonlinear operation unit 28 is passed through a lowpass filtering and downsampling unit 30 to reduce the data rate and consequently reduce the computational requirements of later components of the system. Lowpass filtering and downsampling unit 30 uses an FIR filter computed every other sample for a downsampling factor of two.

A windowing and FFT unit 32 multiplies the output of lowpass filtering and downsampling unit 30 by a window and computes a real input FFT, $S^i(\omega)$, of the product. Typically, windowing and FFT unit 32 uses a Hamming window and a real input FFT.

Finally, a second nonlinear operation unit 34 performs a nonlinear operation on $S^i(\omega)$ to facilitate estimation of voiced or total energy and to ensure that the outputs of channel processing units 20, $T^i(\omega)$, combine constructively if used in fundamental frequency estimation. The absolute value squared is used because it makes all components of $T^i(\omega)$ real and positive.

With reference to Fig. 4, second parameter estimator 16 produces the second preliminary V/UV estimates using a sinusoid detector/estimator. Channel processing units 36 in second parameter estimator 16 divide speech signal s (n) into at least two frequency bands and process the frequency bands to produce a first set of signals, designated as $R^0(I)$ .. $R^I(I)$. Channel processing units 36 are differentiated by the parameters of a bandpass filter used in the first stage of each channel processing unit 36. In the described embodiment, there are sixteen channel processing units (I equals 15). The number of channels (value of I) in Fig. 4 does not have to equal the number of channels (value of I) in Fig. 2.

A remap unit 38 transforms the first set of signals to produce a second set of signals, designated as $S^0(I)$ .. $S^K(I)$. The remap unit can be an identity system. In the described embodiment, there are eight signals in the second set of signals (K equals 7). Thus, remap unit 38 maps the signals from the sixteen channel processing units 36 into eight signals. Remap unit 38 does so by combining consecutive pairs of signals from the first set into single signals in the

second set. For example, $R^0(l)$ and $R^1(l)$ are combined to produce $S^0(l)$, and $R^{14}(l)$ and $R^{15}(l)$ are combined to produce $S^7(l)$. Other approaches to remapping could also be used.

Next, V/UV parameter estimation units 40, each associated with a signal from the second set, produce preliminary V/UV parameters B^0 to B^K by computing a ratio of the sinusoidal energy in the signal to the total energy in the signal and subtracting this ratio from 1:

$$B^k = 1.0 - S^k(1)/S^k(0).$$

With reference to Fig. 5. when speech signal s(n) enters a channel processing unit 36, components $s^i(n)$ belonging to a particular frequency band are isolated by a bandpass filter 26 that operates identically to the bandpass filters of channel processing units 20 (see Fig. 3). It should be noted that, to reduce computation requirements, the same bandpass filters may be used in channel processing units 20 and 36, with the outputs of each filter being supplied to a first nonlinear operation unit 28 of a channel processing unit 20 and a window and correlate unit 42 of a channel processing unit 36.

A window and correlate unit 42 then produces two correlation values for the isolated frequency band $s^i(n)$. The first value, $R^i(0)$, provides a measure of the total energy in the frequency band:

$$R^i(0) = [\frac{1}{2}\sum_{n=0}^{N-1}[\,|s^i(n)|^2 + |s^i(n+S)|^2]\,]^2$$

where N is related to the size of the window and typically defines an interval of 20 milliseconds and S is the number of samples by which the bandpass filter shifts the input speech samples. The second value, $R^i(1)$, provides a measure of the sinusoidal energy in the frequency band:

$$R^i(1) = |\sum_{n=0}^{N-1} s^i(n+S)\, s^{*i}(n)|^2.$$

Combination block 18 produces voiced/unvoiced parameters V^0 to V^K by selecting the minimum of a preliminary V/UV parameter from the first set and a function of a preliminary V/UV parameter from the second set. In particular, combination block produces the voiced/unvoiced parameters as:

$$V^k = \min(A^k, f_B(B^k))$$

where

$$f_B(B^k) = B^k + \alpha(k)\,\beta(\omega_o),$$

$$\beta(\omega_o) = 1.0, \text{ when } \omega_o \geq 2\pi/60.0, \text{ or}$$

$$2\pi/(60\omega_o), \text{ when } \omega_o < 2\pi/60.0$$

and $\alpha(k)$ is an increasing function of k. Because a preliminary V/UV parameter having a value close to zero has a higher probability of being correct than a preliminary V/UV parameter having a larger value. the selection of the minimum value results in the selection of the value that is most likely to be correct.

With reference to Fig. 6, in another embodiment, a first parameter estimator 14' produces the first preliminary V/UV estimate using an autocorrelation domain approach. Channel processing units 44 in first parameter estimator 14' divide speech signal s(n) into at least two frequency bands and process the frequency bands to produce a first set of

frequency band signals, designated as $T^0(l)$ .. $T^K(l)$. There are eight channel processing units (K equals 7) and no remapping unit is necessary.

Next, voiced/unvoiced (V/UV) parameter estimation units 46, each associated with a channel processing unit 44, produce preliminary V/UV parameters A^0 to A^K by computing a ratio of the voiced energy in the frequency band at an estimated pitch period n_o to the total energy in the frequency band and subtracting this ratio from 1:

$$A^k = 1.0 - E_v^k(n_o) / E_t^k .$$

The voiced energy in the frequency band is computed as:

$$E_v^k(n_o) = C(n_o)\ T^k(n_o)$$

where

$$C(n_o) = \frac{1}{\sum_{n=0}^{N-1} w(n)\, w(n+n_o)}$$

N is the number of samples in the window and typically has a value of 101, and $C(n_o)$ compensates for the window roll-off as a function of increasing autocorrelation lag. For non-integer values of n_o, the voiced energy at the nearest three values of n are used with a parabolic interpolation method to obtain the voiced energy for n_o. The total energy is determined as the voiced energy for n_o equal to zero.

With reference to Fig. 7, when speech signal s(n) enters a channel processing unit 44, components $s^i(n)$ belonging to a particular frequency band are isolated by a bandpass filter 48. Bandpass filter 48 uses downsampling to reduce computational requirements, and does so without any significant impact on system performance. Bandpass filter 48 can be implemented as a Finite Impulse Response (FIR) or Infinite Impulse Response (IIR) filter, or by using an FFT. A downsampling factor of S is achieved by shifting the input speech samples by S each time the filter outputs are computed.

A nonlinear operation unit 50 then performs a nonlinear operation on the isolated frequency band $s^i(n)$ to emphasize the fundamental frequency of the isolated frequency band $s^i(n)$. For complex values of $s^i(n)$ (i greater than zero), the absolute value, $|s^i(n)|$, is used. For the real value of $s^0(n)$, no nonlinear operation is performed.

The output of nonlinear operation unit 50 is passed through a highpass filter 52, and the output of the highpass filter is passed through an autocorrelation unit 54. A 101 point window is used, and, to reduce computation, the autocorrelation is only computed at a few samples nearest the pitch period.

With reference again to Fig. 4, second parameter estimator 16 may also use other approaches to produce the second voiced/unvoiced estimate. For example, well-known techniques such as using the height of the peak of the cepstrum, using the height of the peak of the autocorrelation of a linear prediction coder residual, MBE model parameter estimation methods, or IMBE (TM) model parameter estimation methods may be used. In addition, with reference again to Fig. 5, window and correlate unit 42 may produce autocorrelation values for the isolated frequency band $s^i(n)$ as:

$$R^i(l) = Re\left[\sum_n s^i(n+l)\, w(n+l)\, s^{*i}(n)\, w(n)\right]$$

where w(n) is the window. With this approach, combination block 18 produces the voiced/unvoiced parameters as:

$$V^k = \min(A^k, B^k).$$

The fundamental frequency may be estimated using a number of approaches. First, with reference to Fig. 8, a fundamental frequency estimation unit 56 includes a combining unit 58 and an estimator 60. Combining unit 58 sums the $T^i(\omega)$ outputs of channel processing units 20 (Fig. 2) to produce $X(\omega)$. In an alternative approach, combining unit 58 could estimate a signal-to-noise ratio (SNR) for the output of each channel processing unit 20 and weigh the various

outputs so that an output with a higher SNR contributes more to $X(\omega)$ than does an output with a lower SNR.

Estimator 60 then estimates the fundamental frequency (ω_o) by selecting a value for ω_o that maximizes $X(\omega_o)$ over an interval from ω_{min} to ω_{max}. Since $X(\omega)$ is only available at discrete samples of ω, parabolic interpolation of $X(\omega_o)$ near ω_o is used to improve accuracy of the estimate. Estimator 60 further improves the accuracy of the fundamental estimate by combining parabolic estimates near the peaks of the N harmonics of ω_o within the bandwidth of $X(\omega)$.

Once an estimate of the fundamental frequency is determined, the voiced energy $E^v(\omega_o)$ is computed as:

$$E^v(\omega_o) = \sum_{n=1}^{N} \sum_{\omega_m \in I_n} X(\omega_m)$$

where

$$I_n = [(n-0.25)\omega_o, (n+0.25)\omega_o].$$

Thereafter, the voiced energy $E^v(0.5\omega_o)$ is computed and compared to $E^v(\omega_o)$ to select between ω_o and $0.5\omega_o$ as the final estimate of the fundamental frequency.

With reference to Fig. 9, an alternative fundamental frequency estimation unit 62 includes a nonlinear operation unit 64, a windowing and Fast Fourier Transform (FFT) unit 66, and an estimator 68. Nonlinear operation unit 64 performs a nonlinear operation, the absolute value squared, on s(n) to emphasize the fundamental frequency of s(n) and to facilitate determination of the voiced energy when estimating ω_o.

Windowing and FFT unit 66 multiplies the output of nonlinear operation unit 64 to segment it and computes an FFT, $X(\omega)$, of the resulting product. Finally, estimator 68, which works identically to estimator 60, generates an estimate of the fundamental frequency.

With reference to Fig. 10, a hybrid fundamental frequency estimation unit 70 includes a band combination and estimation unit 72, an IMBE estimation unit 74 and an estimate combination unit 76. Band combination and estimation unit 70 combines the outputs of channel processing units 20 (Fig. 2) using simple summation or a signal-to-noise ratio (SNR) weighting where bands with higher SNRs are given higher weight in the combination. From the combined signal ($U(\omega)$), unit 72 estimates a fundamental frequency and a probability that the fundamental frequency is correct. Unit 72 estimates the fundamental frequency by choosing the frequency that maximizes the voiced energy ($E_v(\omega_o)$) from the combined signal, which is determined as:

$$E_v(\omega_o) = \sum_{n=1}^{N} \sum_{\omega_m \in I_n} U(\omega_m)$$

where

$$I_n = [(n-0.25)\omega_o, (n+0.25)\omega_o].$$

and N is the number of harmonics of the fundamental frequency. The probability that ω_o is correct is estimated by comparing $E_v(\omega_o)$ to the total energy E_t, which is computed as:

$$E_t = \sum_{\forall \omega_m > 0.5\omega_o} U(\omega_m) .$$

When $E_v(\omega_o)$ is close to E_t, the probability estimate is near one. When $E_v(\omega_o)$ is close to one half of E_t, the probability estimate is near zero.

IMBE estimation unit 74 uses the well known IMBE technique, or a similar technique, to produce a second fundamental frequency estimate and probability of correctness. Thereafter, estimate combination unit 76 combines the two

fundamental frequency estimates to produce the final fundamental frequency estimate. The probabilities of correctness are used so that the estimate with higher probability of correctness is selected or given the most weight.

With reference to Fig. 11, a voiced/unvoiced parameter smoothing unit 78 performs a smoothing operation to remove voicing errors that might result from rapid transitions in the speech signal. Unit 78 produces a smoothed voiced/unvoiced parameter as:

$$v_s^k(n) = 1.0, \text{ when } v^k(n\text{-}1)v^k(n\text{+}1) = 1 \text{ and}$$

$$v^k(n), \text{ otherwise}$$

where the voiced/unvoiced parameters equal zero for unvoiced speech and one for voiced speech. When the voiced/unvoiced parameters have continuous values, with a value near zero corresponding to highly voiced speech, unit 78 produces a smoothed voiced/unvoiced parameter that is smoothed in both the time and frequency domains:

$$v_s^k(n) = \lambda^k(n)\min(v^k(n), \alpha^k(n), \beta^k(n), \gamma^k(n))$$

where

$$\alpha^k(n) = 2v^{k+1}(n), \text{ when } k=0,1,\ldots,K\text{-}1, \text{ or}$$

$$\infty, \text{ when } k=K;$$

$$\beta^k(n) = 2v^{k-1}(n), \text{ when } k=2, 3,\ldots, K, \text{ or}$$

$$\infty, \text{ when } k=0,1;$$

$$\gamma^k(n) = 0.25v^{k-1}(n) + 0.5v^k(n) + 0.25v^{k+1}(n),$$

$$\text{when } k=1, 2, \ldots, K\text{-}1, \text{ or}$$

$$\infty, \text{ when } k=0, K;$$

$$\lambda^k(n) = 0.8, \text{ when } v_s^k(n\text{-}1) < T^k(n\text{-}1) \text{ and}$$

$$|\omega_o(n)\text{-}\omega_o(n\text{-}1)| < 0.25\,|\omega_o(n)|, \text{ or}$$

$$1, \text{ otherwise;}$$

and $T^k(n)$ is a threshold value that is a function of time and frequency.

With reference to Fig. 12, a voiced/unvoiced parameter improvement unit 80 produces improved voiced/unvoiced parameters by comparing the voiced/unvoiced parameter produced when the estimated fundamental frequency equals ω_o to a voiced/unvoiced parameter produced when the estimated fundamental frequency equals one half of ω_o and selecting the parameter having the lowest value. In particular, voiced/unvoiced parameter improvement unit 80 produces improved voiced/unvoiced parameters as:

$$A^k(\omega_o) = \min(A^k(\omega_o), A^k(0.5\omega_o))$$

where

$$A^k(\omega) = 1.0 - E_v^k(\omega)/E_t^k.$$

With reference to Fig. 13, an improved estimate of the fundamental frequency (ω_o) is generated according to a procedure 100. The initial fundamental frequency estimate ($\tilde{\omega}_o$) is generated according to one of the procedures described above and is used in step 101 to generate a set of evaluation frequencies $\tilde{\omega}^k$. The evaluation frequencies are typically chosen to be near the integer submultiples and multiples of $\tilde{\omega}_o$. Thereafter, functions are evaluated at this set of evaluation frequencies (step 102). The functions that are evaluated typically consist of the voiced energy function $E_v(\tilde{\omega}^k)$ and the normalized frame error $E_f(\tilde{\omega}^k)$. The normalized frame error is computed as

$$E_f(\tilde{\omega}^k) = 1.0 - E_v(\tilde{\omega}^k) / E_t(\tilde{\omega}^k) .$$

The final fundamental frequency estimate is then selected (step 103) using the evaluation frequencies, the function values at the evaluation frequencies, the predicted fundamental frequency (described below), the final fundamental frequency estimates from previous frames, and the above function values from previous frames. When these inputs indicate that one evaluation frequency has a much higher probability of being the correct fundamental frequency than the others, then it is chosen. Otherwise, if two evaluation frequencies have similar probability of being correct and the normalized error for the previous frame is relatively low, then the evaluation frequency closest to the final fundamental frequency from the previous frame is chosen. Otherwise, if two evaluation frequencies have similar probability of being correct, then the one closest to the predicted fundamental frequency is chosen. The predicted fundamental frequency for the next frame is generated (step 104) using the final fundamental frequency estimates from the current and previous frames, a delta fundamental frequency, and normalized frame errors computed at the final fundamental frequency estimate for the current frame and previous frames. The delta fundamental frequency is computed from the frame to frame difference in the final fundamental frequency estimate when the normalized frame errors for these frames are relatively low and the percentage change in fundamental frequency is low, otherwise, it is computed from previous values. When the normalized error for the current frame is relatively low, the predicted fundamental for the current frame is set to the final fundamental frequency. The predicted fundamental for the next frame is set to the sum of the predicted fundamental for the current frame and the delta fundamental frequency for the current frame.

Claims

1. A method of analysing a digitized speech signal to determine excitation parameters for the digitized speech signal, preferably as a step in encoding speech, the method comprising dividing the digitized speech signal into one or more frequency band signals; and, preferably at regular intervals of time, performing the further step of: determining a first preliminary excitation parameter using a first method that includes performing a nonlinear operation on at least one of the frequency band signals to produce at least one modified frequency band signal and determining the first preliminary excitation parameter using the at least one modified frequency band signal; determining at least a second preliminary excitation parameter using at least a second method different from the said first method: and using the first and at least a second preliminary excitation parameters to determine an excitation parameter for the digitized speech signal.

2. A method according to Claim 1, wherein at least one of the second methods uses at least one of the frequency band signals without performing the said nonlinear operation.

3. A method according to Claims 1 or 2, wherein the excitation parameter comprises a voiced/unvoiced parameter for at least one frequency band, said parameter preferably having values that vary over a continuous range.

4. A method according to any preceding claim, further comprising determining a fundamental frequency for the digitized speech signal.

5. A method according to Claim 3, wherein the first preliminary excitation parameter comprises a first voiced/unvoiced parameter for the at least one modified frequency band signal, and wherein the first determining step includes determining the first voiced/unvoiced parameter by comparing voiced energy in the modified frequency band signal to total energy in the modified frequency band signal.

6. A method according to Claim 5, wherein the voiced energy in the modified frequency band signal corresponds to the energy associated with an estimated fundamental frequency for the digitized speech signal.

7. A method according to Claim 5, wherein the voiced energy in the modified frequency band signal corresponds to the energy associated with an estimated pitch period for the digitized speech signal.

8. A method according to Claim 5, wherein the second preliminary excitation parameter includes a second voiced/unvoiced parameter for the at least one frequency band signal, and wherein the second determining step includes determining the second voiced/unvoiced parameter by comparing sinusoidal energy in the at least one frequency band signal to total energy in the at least one frequency band signal.

9. A method according to Claim 5, wherein the second preliminary excitation parameter includes a second voiced/unvoiced parameter for the at least one frequency band signal, and wherein the second determining step includes determining the second voiced/unvoiced parameter by autocorrelating the at least one frequency band signal.

10. A method according to any preceding claim, wherein the said using step emphasizes the first preliminary excitation parameter over the second preliminary excitation parameter in determining the excitation parameter for the digitized speech signal when the first preliminary excitation parameter has a higher probability of being correct than does the second preliminary excitation parameter.

11. A method according to any preceding claim, further comprising smoothing the excitation parameter to produce a smoothed excitation parameter.

12. A method of analysing a digitized speech signal to determine excitation parameters for the digitized speech signal, preferably as a step in encoding speech, the method comprising the steps of: determining preliminary excitation parameters from the digitized speech signal; and smoothing the preliminary excitation parameters to produce excitation parameters.

13. A method according to Claim 12, wherein the preliminary excitation parameters include a preliminary voiced/unvoiced parameter for at least one frequency band and the excitation parameters include a voiced/unvoiced parameter for at least one frequency band, which voiced/unvoiced parameter preferably has values that vary over a continuous range.

14. A method according to Claim 13, wherein the excitation parameters include a fundamental frequency.

15. A method according to Claims 13 or 14, wherein the smoothing step makes the voiced/unvoiced parameter more voiced than the preliminary voiced/unvoiced parameter when voiced/unvoiced parameters that are nearby in time and/or frequency are voiced.

16. A method according to Claim 12, wherein the smoothing step is performed as a function of time and/or frequency.

17. A method of analysing a digitized speech signal to determine excitation parameters for the digitized speech signal, preferably as a step in encoding speech, the method comprising the steps of: estimating a fundamental frequency for the digitized speech signal; evaluating a voiced/unvoiced function using the estimated fundamental frequency to produce a first preliminary voiced/unvoiced parameter; evaluating the voiced/unvoiced function at least using one other frequency derived from the estimated fundamental frequency to produce at least one other preliminary voiced/unvoiced parameter; and combining the first and at least one other preliminary voiced/unvoiced parameters to produce a voiced/unvoiced parameter.

18. A method according to Claim 17, wherein the said at least one other frequency is derived from the said estimated fundamental frequency as a multiple or submultiple of the said estimated fundamental frequency.

19. A method according to Claim 17, wherein the combining step includes choosing the first preliminary voiced/unvoiced parameter as the voiced/unvoiced parameter when the first preliminary voiced/unvoiced parameter indicates that the digitized speech signal is more voiced than does the second preliminary voiced/unvoiced parameter.

20. A method of synthesizing speech using excitation parameters, where the excitation parameters are estimated by using a method for determining such parameters according to any preceding claim.

21. A method of analysing a digitized speech signal to determine a fundamental frequency estimate for the digitized speech signal, comprising the steps of: determining a predicted fundamental frequency estimate from previous

fundamental frequency estimates; determining an initial fundamental frequency estimate; evaluating an error function at the initial fundamental frequency estimate to produce a first error function value; evaluating the error function at at least one other frequency derived from the initial fundamental frequency estimate to produce at least one other error function value; selecting a fundamental frequency estimate using the predicted fundamental frequency estimate, the initial fundamental frequency estimate, the first error function value, and the at least one other error function value.

22. A method according to Claim 21, wherein the said at least one other frequency is derived from the said estimated fundamental frequency as a multiple or submultiple of the said estimated fundamental frequency.

23. A method according to Claim 21, wherein the predicted fundamental frequency is determined by adding a delta factor to a previous predicted fundamental frequency, which delta factor is preferably determined from previous first and at least one other error function values, the previous predicted fundamental frequency, and a previous delta factor.

24. A method of synthesizing speech using a fundamental frequency, where the fundamental frequency is estimated using a method according to any of Claims 21, 22 or 23.

25. A system for analysing a digitized speech signal to determine excitation parameters for the digitized speech signal, comprising: means for dividing the digitized speech signal into one or more frequency band signals; means for determining a first preliminary excitation parameter using a first method that includes performing a nonlinear operation on at least one of the frequency band signals to produce at least one modified frequency band signal and determining the first preliminary excitation parameter using the at least one modified frequency band signal; means for determining a second preliminary excitation parameter using a second method that is different from the above said first method; and means for using the first and second preliminary excitation parameters to determine an excitation parameter for the digitized speech signal.

26. A system for analysing a digitized speech signal to determine excitation parameters for the digitized speech signal, comprising: means for determining preliminary excitation parameters from the digitized speech signal; and means for smoothing the preliminary excitation parameters to produce excitation parameters.

27. A system for analysing a digitized speech signal to determine modified excitation parameters for the digitized speech signal, comprising: means for estimating a fundamental frequency for the digitized speech signal; means for evaluating a voiced/unvoiced function using the estimated fundamental frequency to produce a first preliminary voiced/unvoiced parameter; means for evaluating the voiced/unvoiced function using another frequency derived from the estimated fundamental frequency to produce a second preliminary voiced/unvoiced parameter; and means for combining the first and second preliminary voiced/unvoiced parameters to produce a voiced/unvoiced parameter.

28. A system for analysing a digitized speech signal to determine a fundamental frequency estimate for the digitized speech signal, comprising: means for determining a predicted fundamental frequency estimate from previous fundamental frequency estimates; means for determining an initial fundamental frequency estimate; means for evaluating an error function at the initial fundamental frequency estimate to produce a first error function value; means for evaluating the error function at at least one other frequency derived from the initial fundamental frequency estimate to produce a second error function value; and means for selecting a fundamental frequency estimate using the predicted fundamental frequency estimate, the initial fundamental frequency estimate, the first error function value, and the second error function value.

29. A method of analysing a digitized speech signal to determine a voiced/unvoiced function for the digitized speech signal, comprising: dividing the digitized speech signal into at least two frequency band signals; determining a first preliminary voiced/unvoiced function for at least two of the frequency band signals using a first method; determining a second preliminary voiced/unvoiced function for at least two of the frequency band signals using a second method which is different from the above said first method; and using the first and second preliminary excitation parameters to determine a voiced/unvoiced function for at least two of the frequency band signals.

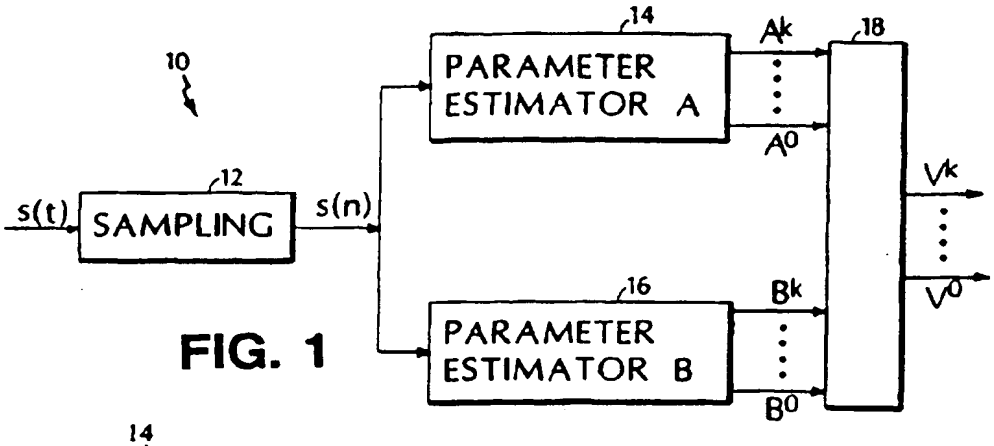

FIG. 1

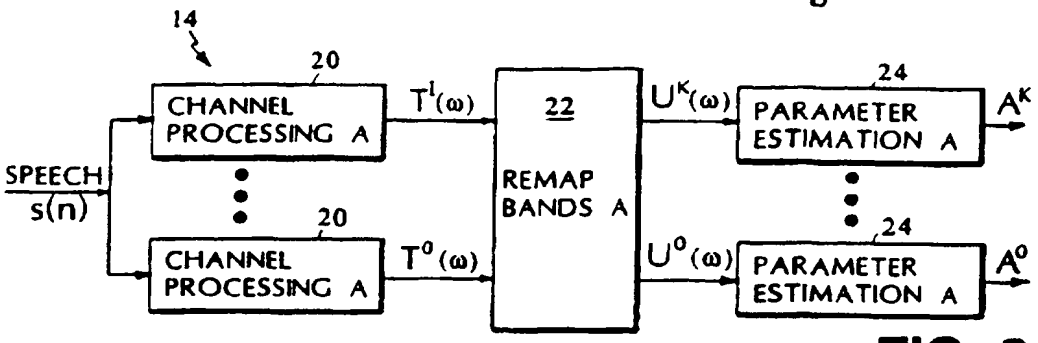

FIG. 2

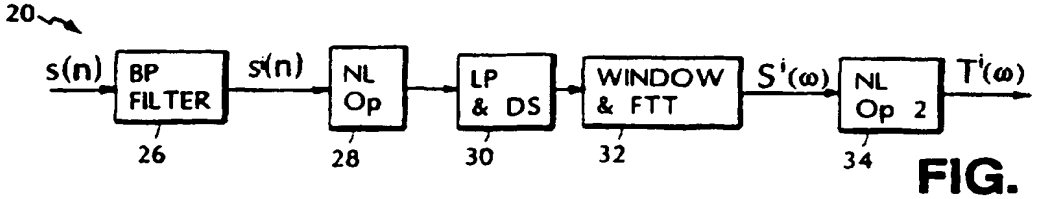

FIG. 3

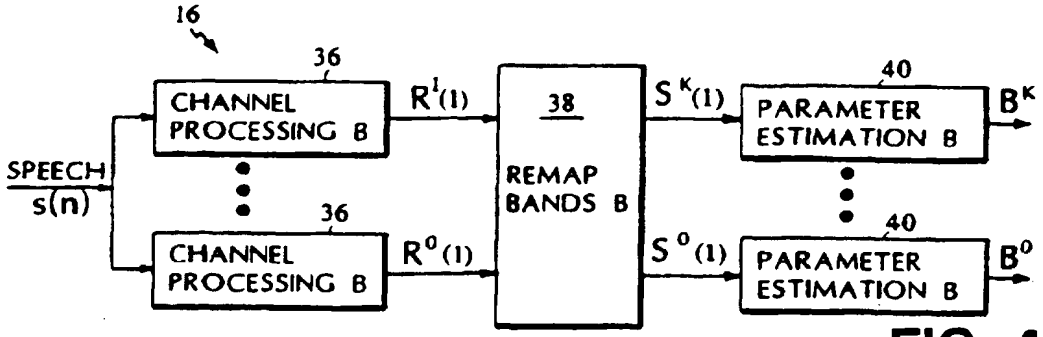

FIG. 4

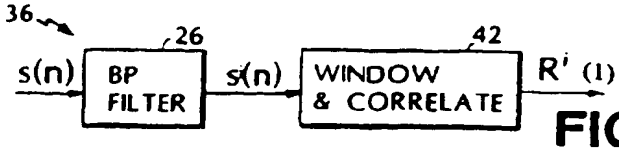

FIG. 5

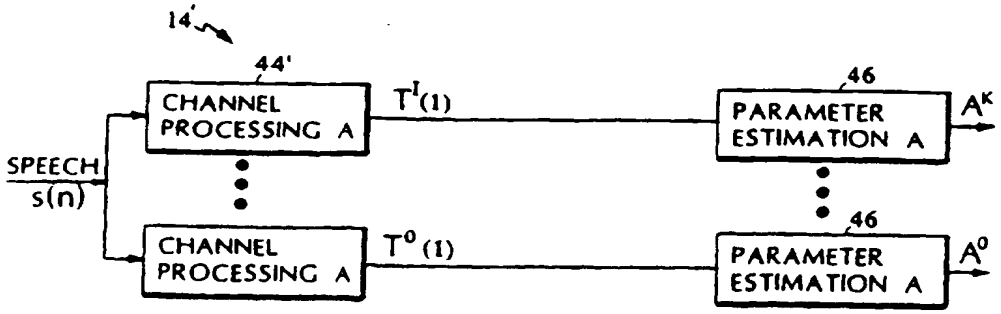

FIG. 6

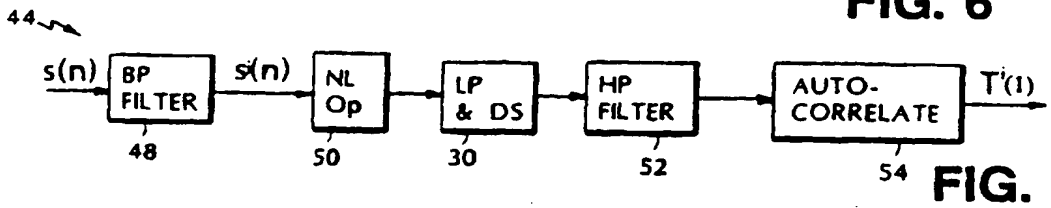

FIG. 7

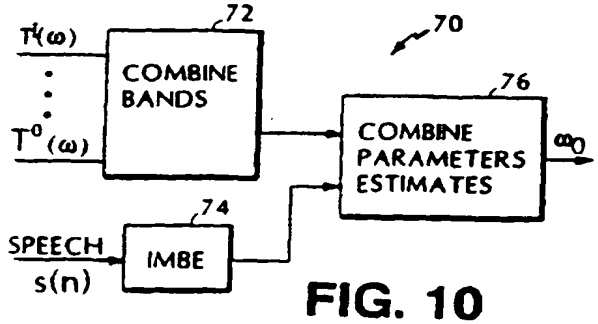

FIG. 10

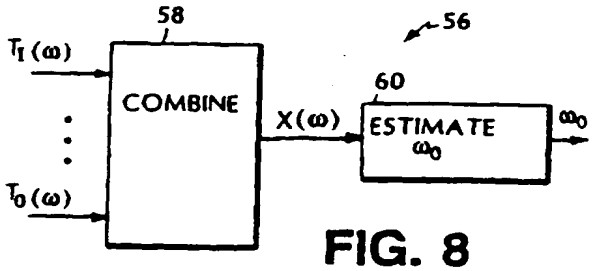

FIG. 8

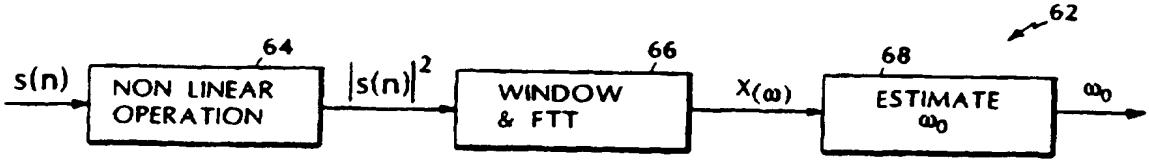

FIG. 9

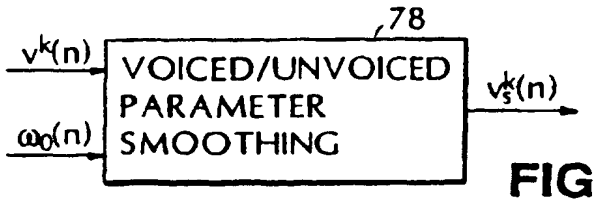

FIG. 11

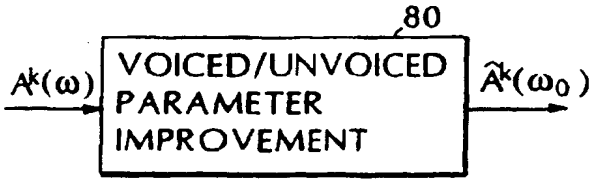

FIG. 12

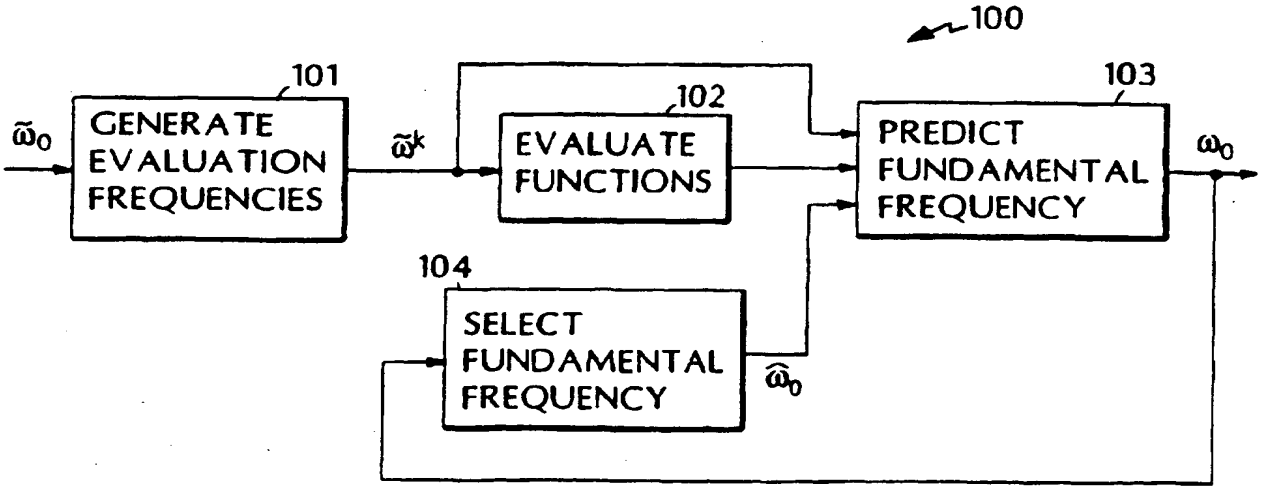

FIG. 13

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 0 722 165 A3

(12) EUROPEAN PATENT APPLICATION

(88) Date of publication A3:
15.07.1998 Bulletin 1998/29

(43) Date of publication A2:
17.07.1996 Bulletin 1996/29

(21) Application number: **96300245.6**

(22) Date of filing: **12.01.1996**

(51) Int Cl.[6]: **G10L 9/14**, G10L 7/04

(84) Designated Contracting States:
DE FR GB SE

(30) Priority: **12.01.1995 US 371743**

(71) Applicant: **DIGITAL VOICE SYSTEMS, INC.**
Burlington,MA 01803 (US)

(72) Inventor: **Griffin, Daniel Wayne**
Hollis, New Hampshire 03049 (US)

(74) Representative: **Deans, Michael John Percy et al**
Lloyd Wise, Tregear & Co.,
Commonwealth House,
1-19 New Oxford Street
London WC1A 1LW (GB)

(54) Estimation of excitation parameters

(57) Excitation parameters for a digitized speech signal are determined by analysing the digitized speech signal. The digitized speech signal is divided into at least two frequency bands. A first preliminary excitation parameter is determined by performing a nonlinear operation on at least one of the frequency band signals to produce a modified frequency band signal and determining the first preliminary excitation parameter using the modified frequency band signal. A second preliminary excitation parameter is determined using a method different from the first method. The first and second preliminary excitation parameters are used to determine an excitation parameter for the digitized speech signal. The method is useful in encoding speech. Speech synthesized using the parameters estimated based on the invention generates high quality speech at various bit rates useful for applications such as satellite voice communication.

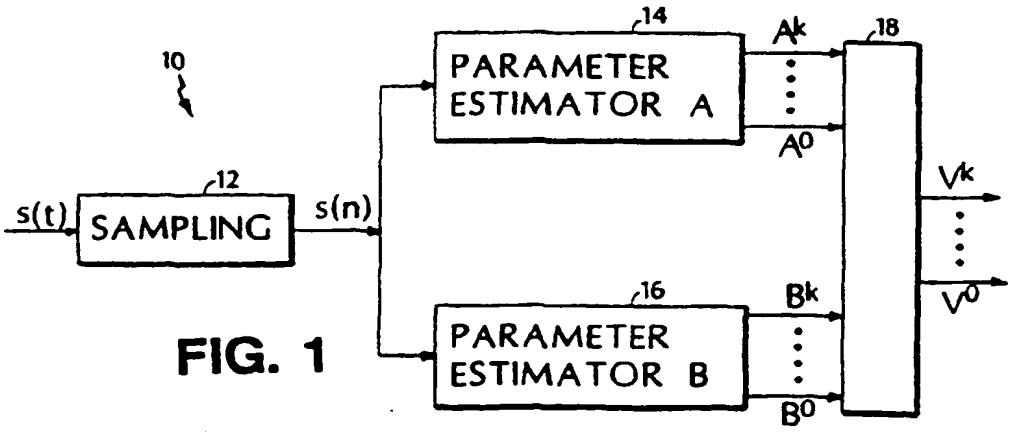

FIG. 1

EP 0 722 165 A3

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 96 30 0245

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
A	DELLER, PROAKIS, HANSEN: "Discrete-time processing of speech signals" 1993 , MACMILLAN PUBLISHING COMPANY XP002051530 * page 460, paragraph 7.4.1 * * page 461; figure 7.25 * ---	1,17,25, 27,29	G10L9/14 G10L7/04
A	WO 88 07740 A (AMERICAN TELEPHONE & TELEGRAPH) * figure 1 * * abstract * ---	1,17,25, 27,29	
A	ICASSP 79. 1979 IEEE INTERNATIONAL CONFERENCE ON ACOUSTICS, SPEECH AND SIGNAL PROCESSING, WASHINGTON, DC, USA, 2-4 APRIL 1979, 1979, NEW YORK, NY, USA, IEEE, USA, pages 69-72, XP002051529 KUREMATSU A ET AL: "A linear predictive vocoder with new pitch extraction and exciting source" * figure 1 * * page 70, column 1, line 3 - line 6 * ---	1,17,25, 27,29	TECHNICAL FIELDS SEARCHED (Int.Cl.6) G10L
A	WO 92 05539 A (DIGITAL VOICE SYSTEMS INC) * page 9, line 1 - line 14 * --- -/--	1,17,25, 27,29	

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	9 January 1998	Krembel, L

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

European Patent Office

CLAIMS INCURRING FEES

EP 96300245.6

The present European patent application comprised at the time of filing more than ten claims.

- ☐ All claims fees have been paid within the prescribed time limit. The present European search report has been drawn up for all claims.
- ☐ Only part of the claims fees have been paid within the prescribed time limit. The present European search report has been drawn up for the first ten claims and for those claims for which claims fees have been paid, namely claims:
- ☐ No claims fees have been paid within the prescribed time limit. The present European search report has been drawn up for the first ten claims.

LACK OF UNITY OF INVENTION

The Search Division considers that the present European patent application does not comply with the requirement of unity of invention and relates to several inventions or groups of inventions, namely:

See Sheet B.

- ☐ All further search fees have been paid within the fixed time limit. The present European search report has been drawn up for all claims.
- ☐ Only part of the further search fees have been paid within the fixed time limit. The present European search report has been drawn up for those parts of the European patent application which relate to the inventions in respect of which search fees have been paid, namely claims:
- ☒ None of the further search fees has been paid within the fixed time limit. The present European search report has been drawn up for those parts of the European patent application which relate to the invention first mentioned in the claims, namely claims: 1-11, 13-20, 25, 27, 29

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 96 30 0245

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.6)
A	IEEE TRANSACTIONS ON SIGNAL PROCESSING, vol. 39, no. 2, 1 February 1991, pages 319-329, XP000206434 KRUBSACK D A ET AL: "AN AUTOCORRELATION PITCH DETECTOR AND VOICING DECISION WITH CONFIDENCE MEASURES DEVELOPED FOR NOISE-CORRUPTED SPEECH" * figure 1 * * Paragraph II.B "Features for the Voicing Decision and Confidence Measures" * * Paragraph II.C "Voicing Decision" * * Paragraph II.D "Voicing Confidence" * -----	10	
			TECHNICAL FIELDS SEARCHED (Int.Cl.6)

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	9 January 1998	Krembel, L

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

European Patent Office

LACK OF UNITY OF INVENTION
SHEET B

Application Number
EP 96 30 0245

The Search Division considers that the present European patent application does not comply with the requirements of unity of invention and relates to several inventions or groups of inventions, namely:

1. Claims: 1-11,17-20,25,27,29

   Determination of excitation parameters using a combination of two preliminary excitation parameters.

2. Claims: 12-16,26

   Smoothing of excitation parameters.

3. Claims: 21-24,28

   Fundamental frequency tracking method.